C&T Riqchary
Revista de Investigación en Ciencia y Tecnología
ISSN: 2810-8124 (en línea) / ISSN: 2706-543x

Vol. 6 Núm. 1 (2024)
Publicado: 22/03/24
https://doi.org/10.57166/riqchary.v6.n1.2024.118

Estudio comparativo entre técnicas estadísticas multivariadas y redes neuronales artificiales para la optimización de la vigilancia de la calidad de agua para consumo humano en la red de salud Abancay 2022

Comparative study between multivariate statistical techniques and artificial neural networks for the optimization of the surveillance of water quality for human consumption in the Abancay health network 2022

Frank Michael Zuloaga Estacio[A], Mario Aquino Cruz[B]

ORCID:0009-0003-8946-5030[A], ORCID:0000-0002-2552-5669[B]

(Recepción: 13/02/2024 y aceptación 10/03/2024)

Resumen— Actualmente en la mayoría de instituciones, incluyendo la red de salud Abancay, se utiliza estadística tradicional con el fin de determinar tendencias sobre un conjunto de datos que se centra en una sola variable, es complicado aplicar este tipo de análisis a conjunto de datos multivariados, que son los que usualmente se obtienen en los programas de control de calidad de agua y, que excluye diferencias entre las variables analizadas y sus relaciones. El objetivo del estudio es realizar una comparación entre las diferentes técnicas estadísticas multivariables y redes neuronales artificiales, con la finalidad de relacionar y clasificar las variables. Para realizar esto se eligió dos técnicas de estadística multivariada, análisis de componentes principales (ACP) y análisis discriminante (AD) y dos tipos de redes Neuronales artificiales, de aprendizaje no supervisado, hebbiano (RNAH), y de aprendizaje supervisado, perceptrón multicapa (RNAPM), el tipo de investigación que se usará en el estudio será investigación aplicada de enfoque cuantitativo, con un nivel de investigación explicativo y con un diseño transversal. Dada la comparación entre el análisis de componentes principales y la red neuronal artificial de tipo Hebbiano, se obtuvo que las redes neuronales pudieron asociar mejor las variables que el análisis de componentes principales. En la segunda comparación entre el análisis discriminante y la red neuronal artificial perceptrón multicapa los resultados fueron buenos para el análisis discriminante debido a que obtuvo un 95% de clasificación correcta, mientras que la red neuronal artificial obtuvo un 74.6%, sin embargo, debido a las limitaciones del análisis discriminante, se infiere que la red neuronal artificial perceptrón multicapa es un mejor modelo a escoger.

Palabra clave: análisis de componentes principales, análisis discriminante, red neuronal artificial de tipo hebbiano, red neuronal artificial perceptrón multicapa.

Abstract— Currently, in most institutions, including the Abancay health network, traditional statistics are used in order to determine trends on a data set that focuses on a single variable, it is complicated to apply this type of analysis to multivariate data sets, which are those usually obtained in water quality control programs, and which exclude differences between the variables analyzed and their relationships. The objective of the study was to make a comparison between different multivariate statistical techniques and artificial neural networks, in order to relate and classify variables. To do this, two multivariate statistical techniques were chosen, principal component analysis (PCA) and discriminant analysis (DA) and two types of artificial neural networks, unsupervised learning, hebbian (RNAH), and supervised learning, multilayer perceptron (RNAPM), the type of research I used in the study will be applied research of quantitative approach, with an explanatory level of research and with a cross-sectional design. Given the comparison between the principal component analysis and the Hebbian type artificial neural network, I obtained that the neural networks were able to associate the variables better than the principal component analysis. In the second comparison between the discriminant analysis and the multilayer perceptron artificial neural network, the results were good for the discriminant analysis because it obtained 95% of correct classification, while the artificial neural network obtained 74.6%, however, due to the limitations of the discriminant analysis, I inferred that the multilayer perceptron artificial neural network is a better model to choose.

Keyword: discriminant analysis, hebbian-type artificial neural network, multilayer perceptron artificial neural network, principal component analysis.

A. *Frank Michael Zuloaga Estacio[A], Escuela Profesional de Ingeniería Informática y Sistemas de la Universidad Nacional Micaela Bastidas de Apurímac-Perú, frankzuloaga3@gmail.com*

B. *Mario Aquino Cruz[B] Escuela Profesional de Ingeniería Informática y Sistemas de la Universidad Nacional Micaela Bastidas de Apurímac-Perú, maquino@unamba.edu.pe*

1 INTRODUCCIÓN

El análisis de componentes principales es una técnica estadística general que reduce una tabla de datos variables a sus características básicas, llamadas componentes principales. Los componentes principales son ciertas combinaciones lineales de las variables originales que explican la varianza máxima de todas las variables. [1]. En el proceso, el método proporciona una aproximación de la tabla de datos original utilizando solo estos pocos componentes principales.

Su objetivo es extraer información importante de datos estadísticos y representarla como un nuevo conjunto de variables ortogonales llamadas componentes principales y mostrar el patrón de similitudes entre observaciones y variables en forma de puntos en un mapa de puntos. Matemáticamente, está determinado por vectores propios y valores propios. Los vectores propios y los valores propios son números y vectores asociados con matrices cuadradas. Juntos proporcionan una descomposición matricial adecuada para analizar la estructura de esa matriz, como una matriz de correlación, covarianza o producto. [2]. Si bien, se han listado diferentes tipos de matrices para el estudio se eligió la matriz de correlación.

El análisis de función discriminante es un análisis estadístico que se utiliza para analizar datos cuando la variable dependiente o el resultado es categórico y la variable independiente o predictor es paramétrico. Este es un método paramétrico para determinar qué pesos de variables continuas o predictores diferencian mejor dos o más tipos de variables dependientes y lo hacen mejor que el azar. El análisis discriminante se utiliza para determinar la precisión de un sistema de clasificación determinado al predecir una muestra dentro de un grupo en particular. El análisis discriminante implica desarrollar funciones discriminantes para cada muestra y obtener una puntuación umbral utilizada para clasificar la muestra en diferentes grupos. [3].

Los algoritmos de aprendizaje no supervisado de carácter hebbiano se basan en el siguiente postulado, formulado por Donald O. Hebb en 1949: "Cuando un axón de una celda A esta suficientemente cerca para conseguir excitar una celda B y repetida o persistentemente toma parte en su activación, algún proceso de crecimiento o cambio metabólico tiene lugar en una o ambas celdas, de tal forma que la eficiencia de A, cuando la celta activa es B, aumenta". De esta forma, identificando las celdas con neuronas fuertemente conectadas [4]. La regla de Hebb es de tipo no supervisado, pues la modificación de los pesos depende de los estados (salidas) de las neuronas obtenidos tras la presentación de un estímulo determinado, con independencia de que coincidan o no con las deseadas. De esta forma, en el aprendizaje hebbiano múltiples neuronas de salida pueden activarse simultáneamente.

La arquitectura de un perceptrón multicapa es variable, pero en general consiste en varias capas de neuronas, puede tener una o más capas ocultas y finalmente una capa de salida. El perceptrón multicapa se ha aplicado a una amplia variedad de tareas, todas las cuales pueden clasificarse como predicción, aproximación de funciones o clasificación de patrones. La predicción implica pronosticar tendencias futuras en una serie temporal de datos dadas las condiciones actuales y anteriores. La aproximación de funciones se ocupa de modelar la relación entre variables. La clasificación de patrones implica clasificar datos en clases discretas. Todas estas aplicaciones están estrechamente relacionadas [5].

Este artículo presenta los resultados de la aplicación y comparación entre el análisis de componentes principales con la red neuronal de tipo hebbiano y el análisis discriminante con una red neuronal perceptrón multicapa. Los datos analizados corresponden a las variables de calidad de agua correspondientes a cloro, pH, conductividad, turbiedad, debido a que solo estas estaban presentes en los datos obtenidos, además se añadió una variable periodo para realizar la comparación entre el análisis discriminante y la red neuronal perceptrón multicapa debido a que la predicción de datos no se puede realizar si no existe una variable de salida, la cual se obtuvo al analizar los datos de precipitación en la zona durante el mes de enero a diciembre desde 1991 al 2021.

2 METODOLOGÍA

El tipo de investigación que se realiza en el presente estudio es una investigación aplicativa con un enfoque cuantitativo, el nivel de investigación será explicativo y su diseño es transversal.

2.1 ASOCIACIÓN DE VARIABLES

Para determinar la asociación entre las variables fisicoquímicas que hacen parte de los datos que se adquieren comúnmente en los programas de monitoreo de calidad de agua se emplearon técnicas de análisis de componentes principales y las redes neuronales artificiales de tipo Hebbiano.

2.1.1 ANÁLISIS DE COMPONENTES PRINCIPALES

Inicialmente se elaboró histogramas para poder observar si la distribución presentada por las variables se trataba de una distribución simétrica o normal o si se requería el uso de ciertos parámetros para evitar que los resultados se distorsionen por ejemplo haciendo uso de la transformación

logarítmica para disminuir su rango y posteriormente realizar una estandarización de los datos como paso preliminar para la determinación de los componentes principales.

En este caso se observó que ninguna de las variables tenía una distribución normal (cloro, conductividad, pH, turbiedad)

Como se puede observar en el valor de significancia ninguna de las variables presento una distribución normal usando el método de Kolmogorov-Smirnov debido a la cantidad de datos, por lo cual se realizó transformación logarítmica a todas las variables para evitar que los datos distorsionen los resultados del análisis estadístico.

Realizada la transformación logarítmica. Se procedo a estandarizar los datos previos a la determinación de los componentes principales (CP). El análisis de CP arrojo como resultados un conjunto de vectores propios que corresponden a un nuevo espacio en donde se proyectan los datos originales. Con estos, se pudo calcular las coordenadas de los datos en este nuevo espacio, para efecto de visualización y para determinar asociaciones de variables presentes en los datos. Esto último se pudo estudiar mediante la matriz de coeficientes de correlación entre los CP y las variables originales.

En la figura 1. podemos observar que no existe una pendiente aparente por lo cual se pueden utilizar todos los componentes para el estudio.

Fig. 1. Gráfico de Sedimentación de los componentes principales

Así mismo, con el fin de determinar las asociaciones entre las variables fisicoquímicas estudiadas se calculó la matriz de coeficientes de correlación entre los CP a utilizar en el estudio y las variables originales.

2.1.2 Red neuronal artificial de tipo hebbiano

Todo el procedimiento computacional de las redes se realizó con las funciones diseñadas en Matlab.

El tipo de red neuronal utilizada fue la red neuronal de tipo hebbiano con entrenamiento hebbiano generalizado (aprendizaje no supervisado)

El procedimiento seguido para determinar los componentes principales con RNAH fue el siguiente.

1. Estandarización de los datos en SPSS
2. Cargar los datos estandarizados en Matlab.
3. Se realizo el entrenamiento de red con los siguientes parámetros.
4. 4 variables de entrada
5. 4 variables de salida
6. La tasa de aprendizaje fue igual a $1x10^{-6}$
7. Numero de épocas de entrenamiento igual a 1000

A esta red no se le determino el coeficiente de correlación, puesto que es una red de entrenamiento no supervisado, y no se contó con una salida deseada con la cual se pudiera comparar.

2.2 Clasificación de variables

Para realizar este tipo de análisis usando las funciones discriminantes y las RNAPM se hizo necesario etiquetar las muestras para que los métodos aplicados pudiesen predecir la pertenencia de una variable o parámetro a un determinado grupo. Los grupos establecidos fueron dos: seco y húmedo, de acuerdo a la fecha de toma de las muestras analizadas. Esto se hizo con la finalidad de establecer si el periodo hidrológico afectaba de forma significativa los valores de las concentraciones de los parámetros analizados.

Para realizar el ejercicio de clasificación, no se utilizó ninguna técnica específica para definir los grupos, el investigador lo determino según el análisis de la información de precipitación del área de estudio, como herramienta de apoyo en la selección de periodos secos y húmedos se graficó la precipitación promedio mensual, en las estaciones ubicadas en el embalse, así como otras estaciones pluviométricas ubicadas en los alrededores de la zona de estudio. A los datos de precipitación se les realizo un tratamiento estadístico para su análisis, ya que en este caso particular lo que interesaba era el patrón de variación temporal y la determinación de la ocurrencia de periodos secos y húmedos. Esta parte del análisis se encuentra motivada por la variabilidad que han presentado los parámetros hidrológicos en los últimos años. A continuación, en la tabla 1, se presenta el histograma de la precipitación promedio mensual entre 1991 y 2021.

TABLA 1
Tabla de climatología

Mes	Medida		
	Temperatura media (°C)	Precipitación (mm)	Días lluviosos (días)
Enero	10.5	225	21
Febrero	10.4	205	19

Marzo	10.3	185	21
Abril	10	100	17
Mayo	9.5	35	8
Junio	8.7	18	3
Julio	8.4	18	4
Agosto	9.2	29	6
Septiembre	10	50	12
Octubre	10.6	108	18
Noviembre	11.1	132	18
Diciembre	10.7	185	20

Como base a un análisis visual de la tabla, se determinó, por criterio del investigador, que todos los datos por debajo de 60 mm de precipitación serian periodos secos y por encima de estos serían periodos húmedos.

2.2.1 Análisis discriminante

Para poder correr la función discriminante fue necesario conseguir que los datos fueran de distribución normal, para lograr eso se realizó transformación de distribución inversa sobre los datos de una de las estaciones con la intención de utilizar el modelo en una muestra más amplia para comprobar su validez. Sin embargo, debido a la cantidad de datos anómalos en diferentes variables, se decidió utilizar solo una variable, la variable de precipitación la cual permitirá que se cumpla la prueba de homogeneidad de matrices de covarianza M de Box cuyo valor de significancia fue de 0.589 la cual prueba que el método se puede usar sobre los datos.

Al correr la función, se determinaron los puntos medios y un centroide a cada periodo, cuyos valores son presentados en la tabla 2.

TABLA 2
Funciones de centroide

Variable	Centroide
Periodo Húmedo	1.316
Periodo Seco	-2.057

Esto significa que el punto medio donde se definirá si un elemento es del periodo húmedo o seco será en el valor -0.3705, teniendo en cuenta esto si una muestra tiene un valor de coordenada menor a -0.3705 se clasificara como periodo seco, y en caso contrario como periodo húmedo.

2.2.1 Red neuronal artificial perceptrón multicapa

Para realizar el Análisis discriminante se utilizó una RNA de entrenamiento supervisado con retropropagación de error y alimentación hacia adelante (perceptrón multicapa). Esta red se seleccionó debido a que es la red más sencilla para hacer un análisis equivalente al análisis discriminante.

La cantidad de datos escogidos es el total donde se definió que 571 datos se encontraban en el periodo húmedo y 400 en el periodo seco los cuales son mostrados en la tabla 3.

TABLA 3
Datos escogidos

	Precipitación			
Periodos	Frecuencia	Porcentaje	Porcentaje Válido	Porcentaje Acumulado
Periodo Húmedo	571	58.8	58.8	58.8
Periodo Seco	400	41.2	41.2	100.0
Total	971	100.0	100.0	

Los pasos que se realizaron para el diseño del modelo de red neuronal han sido los siguientes:

Selección de variables, procesamiento de datos, creación de grupos y selección y construcción del modelo de red neuronal.

1. Selección de variables: las variables seleccionadas para confeccionar la muestra fueron: cloro, conductividad, pH y turbiedad. Son 4 variables en total, con lo que se usaran 4 neuronas de entrada y 1 de salida, ya que se busca si la muestra se tomó en periodo húmedo o seco. La función usada en las neuronas de la capa oculta es una función tangente hiperbólica. En cambio, en la neurona de salida se aplicará una función sigmoidal, ya que se realiza una tarea de predicción y para estas tareas se recomienda el uso de esta función, ya que los valores no son continuos, sino que dentro de los valores dados (0 y 1), la respuesta se puede desplazar de forma no continua en ellos. En la capa oculta se usará una función tangente hiperbólica, ya que, al implementarla en una red neuronal, se pueden lograr salidas simétricas y balanceadas, lo que permite modelar relaciones no lineales entre los datos con mayor precisión.
2. Procesamiento de datos: Las 4 variables que se han tenido en cuenta a la hora de conformar la muestra se transformaron y estandarizaron para establecerla en una nueva distribución comprendida.
3. Creación de los grupos de entrenamiento, validación y reserva: La muestra de 971 datos se han dividido en tres grupos: entrenamiento, el cual se usará en la fase de aprendizaje, validación, usado en la fase de funcionamiento, aplicado en la etapa para demostrar el funcionamiento de la red. Para ello, el grupo de entrenamiento

consta del 61.5% de los datos para que la red sea capaz de ajustar los pesos de forma iterativa. Luego el grupo de validación estará compuesto por el 27.6% y el grupo de reserva está compuesto por el 10.9% de los datos.

4. Selección y construcción del modelo de red neuronal: Se usará el modelo de red de perceptrón multicapa, ya que es la red de interés del trabajo y que además se ajusta al objetivo del mismo, el cual es la clasificación de un grupo de muestras para saber si se recogieron en un periodo húmedo o seco. Para encontrar el modelo más eficaz se creó diferentes modelos modificando los siguientes valores: número de neuronas ocultas y tasa de aprendizaje. El primer valor a modificarse ha sido el número de neuronas ocultas para encontrar el modelo con el mejor desempeño, el número de neuronas que se han aplicado han sido 1, 2, 3, 4 y 5. Después se ha modificado el valor de la tasa de aprendizaje a los valores 0.4, 0.2, 0.1, 0.05, sin embargo, se decidió no incluir un cambio en el número de capas ocultas debido a que dichos modelos no mostraron que él porcentaje de acierto varíe sustancialmente. Para ello se han creado 20 modelos de red neuronal, se observó el modelo que tuvo un desempeño más equilibrado y mayor en las fases de entrenamiento y validación, para que su desempeño en la fase de test sea más estable. Siguiendo estos criterios el modelo que se seleccionó para pasar a la fase de test fue el modelo número 17 debido a que aparte de ser el modelo con mejor porcentaje de acierto de clasificación en la fase de entrenamiento no existe mucha diferencia con el porcentaje de acierto de la fase de validación. Una vez seleccionado el modelo, se pasó a la fase final, la fase test, para obtener el rendimiento real de la red, la arquitectura del modelo 17 es representado gráficamente en la figura 2.

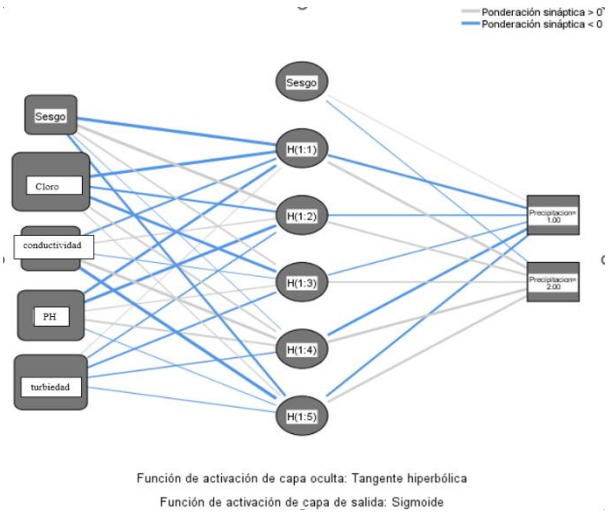

Fig. 2. Arquitectura del modelo de red neuronal escogido.

5. Evaluación de rendimiento del modelo: El análisis se hizo a través de la curva de ROC obtenida a través del programa SPSS. Se usó este método de análisis, ya que nos encontramos ante un problema de clasificación de dos variables y este es la medida precisa y válida para evaluar la precisión diagnostica del modelo seleccionado.

3 Resultados y discusión

3.1 Resultados del análisis de componentes principales

En la siguiente tabla se puede observar la relación entre los componentes principales escogidos para el estudio y las variables originales, para ello se utilizó el método de Spearman debido a que la distribución no es normal.

TABLA 4
Matriz de correlaciones ACP

Variable	Correlaciones			
	Componente 1	Componente 2	Componente 3	Componente 4
Cloro	0.494	-0.849	0.702	0.353
Conductividad	0.744	0.300	-0.501	0.307
pH	0.131	0.569	-0.008	-0.039
Turbiedad	-0.797	0.124	0.007	0.468

Siguiendo el análisis de la matriz de correlación, se observó que la mayoría de los valores de dichos coeficientes resultaron tener valores suficientemente altos. Para efectos del presente trabajo se considera que la relación entre dos variables es significativa si es mayor a 0.4 en valor absoluto.

En la tabla 4 se muestra la correlación entre las variables fisicoquímicas identificadas a partir del Análisis de Componentes Principales. Como se evidencia de la inspección de dicha tabla, existe muy poca relación entre el componente 1 y la variable pH mientras que el componente 2 tiene muy poca relación con la variable conductividad y turbiedad, el componente 3 tiene muy poca relación con pH y turbiedad mientras que el componente 4 tiene muy poca relación con pH y tiene apreciable relación con cloro y conductividad Sin embargo, todas las variables se relacional al menos con un componente por lo cual se puede decir que existe una correlación considerable entre las variables analizadas en el presente estudio.

3.2 Resultados de la red neuronal artificial de tipo hebbiano

En la tabla 5, se puede observar que existe asociación significativa entre las variables estudiadas y los componentes principales obtenidos mediante la red neuronal artificial Hebbiana.

TABLA 5
Matriz de correlaciones RNAH

Variable	Correlaciones			
	Componente 1	Componente 2	Componente 3	Componente 4
Cloro	-0.258	0.258	0.258	0.775
Conductividad	0.600	0.800	0.400	0.400
pH	0.000	0.400	0.800	0.800
Turbiedad	-0.200	-0.400	-0.800	0.000

De esta se puede ver que el componente 1 agrupa a las variables Conductividad, el componente 2 agrupa a todas las variables menos Cloro, la tercera agrupa a todas las variables menos Cloro, la cuarta a todas menos a turbiedad.

De lo anterior se infiere que en los componentes 1 no se encuentra relación significativa entre las variables originales, sin embargo, en el componente 2 y 3 se agrupa 2 variables químicas y una física y en el componente 4 se agrupan 3 variables químicas.

3.2 Análisis de resultados de la comparación entre red neuronal artificial de tipo hebbiano y análisis de componentes principales

Como se observó en los resultados de cada metodología, las RNA de tipo hebbiano obtuvieron un mejor desempeño al poder asociar más variables entre sí. Lo que indica que la no linealidad de la RNA permitió encontrar asociaciones entre variables más concretas, que las de la metodología convencional.

Podríamos partir del hecho que la dinámica físico-química que se presenta en el embalse es compleja, y por lo tanto la identificación de los procesos específicos resulta de vital importancia. La metodología de ACP convencional fue menos útil al determinar las asociaciones entre las variables, que son las que permiten identificar procesos fisicoquímicos; mientras que con las RNA de tipo hebbiano las asociaciones se pudieron interpretar de una manera más clara.

3.3 Resultados del análisis discriminante

Los resultados obtenidos por la función se presentan en la tabla 6 donde podemos comparar los datos reales con lo modelado, además se considera que el periodo húmedo tiene el valor de 1 y el periodo seco el valor de 2.

TABLA 6
Matriz de resultado del análisis discriminante

Nombre de la estación	Fecha	Periodo Real	Periodo Modelado
Quisapata alta	22-01-2021 09:34 AM	1	1
Quisapata alta	22-01-2021 09:34 AM	1	1
Quisapata alta	22-01-2021 09:34 AM	1	1
Quisapata alta	28-02-2021 04:03 PM	1	1
Quisapata alta	28-02-2021 04:03 PM	1	1
Quisapata alta	28-02-2021 04:03 PM	1	1
Quisapata alta	28-02-2021 04:03 PM	1	1
Quisapata alta	22-03-2021 12:34 PM	1	1
Quisapata alta	22-03-2021 12:34 PM	1	1
Quisapata alta	22-03-2021 12:34 PM	1	1
Quisapata alta	22-03-2021 12:34 PM	1	1
Quisapata alta	24-05-2021 10:43 AM	2	2
Quisapata alta	24-05-2021 10:43 AM	2	2
Quisapata alta	24-05-2021 10:43 AM	2	2
Quisapata alta	24-05-2021 10:43 AM	2	2
Quisapata alta	26-06-2021 08:26 PM	2	2
Quisapata alta	26-06-2021 08:26 PM	2	2
Quisapata alta	26-06-2021 08:26 PM	2	2
Quisapata alta	26-06-2021 08:26 PM	2	2
Quisapata alta	25-07-2021 08:33 PM	2	2
Quisapata alta	25-07-2021 08:33 PM	2	2
Quisapata alta	25-07-2021 08:33 PM	2	2
Quisapata alta	25-07-2021 08:33 PM	2	2
Quisapata alta	24-09-2021 11:55 AM	2	2
Quisapata alta	24-09-2021 11:55 AM	2	2
Quisapata alta	24-09-2021 11:55 AM	2	2
Quisapata alta	24-09-2021 11:55 AM	2	2
Quisapata alta	31-10-2021 06:35 PM	1	1
Quisapata alta	31-10-2021 06:35 PM	1	1
Quisapata alta	31-10-2021 06:35 PM	1	1
Quisapata alta	31-10-2021 06:35 PM	1	1
Quisapata alta	21-11-2021 04:54 PM	1	1
Quisapata alta	21-11-2021 04:54 PM	1	1
Quisapata alta	21-11-2021 04:54 PM	1	1
Quisapata alta	21-11-2021 04:54 PM	1	1
Quisapata alta	24-11-2021 11:09 AM	1	1
Quisapata alta	15-12-2021 04:04 PM	1	1
Quisapata alta	15-12-2021 04:04 PM	1	1
Quisapata alta	15-12-2021 04:04 PM	1	1
Quisapata alta	15-12-2021 04:04 PM	1	1

A simple vista se puede observar que el periodo modelado y el periodo real fueron clasificados correctamente en el 100% de los casos, esto lo podemos ver en la tabla 7 de validación cruzada.

TABLA 7
Tabla de validación cruzada análisis discriminante

Validación Cruzada		Pertenencia a grupos pronosticada		
	Precipitación	Periodo Húmedo	Periodo Seco	Total
Recuento	Periodo Húmedo	25	0	25
	Periodo Seco	0	16	16
%	Periodo Húmedo	100.0	0.0	100.0
	Periodo Seco	0.0	100.0	100.0

100% de los casos agrupados fueron clasificados correctamente.

Si probamos este modelo con todos los datos el valor de significancia que obtendremos en la prueba de homogeneidad m de box será de 0 debido a que existe algún valor que muestre disparidad, sin embargo, podríamos calcular un estimado de lo que sería una validación cruzada con todos los datos disponibles para probar que el modelo funciona, lo cual se muestra en la tabla 8.

TABLA 8
Tabla de validación cruzada análisis discriminante con todos los datos

Validación Cruzada		Pertenencia a grupos pronosticada		
	Precipitación	Periodo Húmedo	Periodo Seco	Total
Recuento	Periodo Húmedo	523	48	571
	Periodo Seco	0	400	400
%	Periodo Húmedo	91.6	8.4	100.0
	Periodo Seco	0.0	100.0	100.0

Como se puede observar que el 91.6% de los datos fueron clasificados correctamente en el Periodo húmedo y el 100% en el Periodo seco, teniendo esto en cuenta podemos decir que el 95.1% de los casos fueron clasificados correctamente.

Estos resultados indican que el análisis discriminante convencional puede realizar una clasificación satisfactoria del periodo hidrológico a partir de los valores de las muestras.

Sin embargo, como fue mencionado anteriormente el modelo es inestable si se ejecuta en una cierta cantidad de datos debido a que la significancia del M de Box será 0.

3.3 Resultados de la red neuronal artificial perceptrón multicapa

Teniendo en cuenta el modelo ya escogido los resultados son los siguientes:

1. Grupo de entrenamiento: el grupo de entrenamiento, formado por 597 sujetos se obtuvo que 358 datos se encontraban en el periodo húmedo y 239 en el periodo seco. Pero de esta clasificación, la red consiguió clasificar correctamente el 74.5% de los patrones. De los 358 datos en el periodo húmedo 279 son correctos y 79 fueron discriminados incorrectamente. Al igual en los datos de periodo seco, en donde 166 fueron clasificados correctamente y 73 no lo están. Estos datos aparecen en la tabla 9.

TABLA 9
Tabla de clasificación de datos de entrenamiento

Entrenamiento Observado	Pertenencia a grupos pronosticada		
	Periodo Húmedo	Periodo Seco	Porcentaje Correcto
Periodo Húmedo	279	73	79.3%
Periodo Seco	79	166	67.8%
Porcentaje Global	60.0%	40.0%	74.5%

2. Grupo de validación: En cuanto al grupo de validación, formado por 268 datos, de los cuales 157 fueron clasificados como periodo húmedo y 111 como periodo seco. Del grupo de periodo húmedo 116 son verdaderos positivos y 41 falsos positivos. En el grupo de periodo seco 79 fueron clasificados correctamente y 32 no. Por lo tanto, el 72.8% de los patrones se encuentran bien clasificados, como se muestra en la tabla 10.

TABLA 10
Tabla de clasificación de datos de validación

Validación	Pertenencia a grupos pronosticada		
Observado	Periodo Húmedo	Periodo Seco	Porcentaje Correcto
Periodo Húmedo	116	32	78.4%
Periodo Seco	41	79	65,8%
Porcentaje Global	58.6%	41.4%	72.8%

3. Grupo de reserva: En cuanto al grupo de reserva de 106 participantes, el cual es el que nos interesa se muestra el verdadero rendimiento de la red. De este obtenemos que el modelo pudo predecir correctamente 85 (80%) patrones. De los cuales 57 de ellos son del periodo húmedo y 28 del periodo seco. De los 21 patrones mal predichos, 7 son del

periodo húmedo y 14 del periodo seco, la cual es reflejada en la tabla 11.

TABLA 11
Tabla de clasificación de datos de reserva

Reserva	Pertenencia a grupos pronosticada		
Observado	Periodo Húmedo	Periodo Seco	Porcentaje Correcto
Periodo Húmedo	57	14	80.3%
Periodo Seco	7	28	80.0%
Porcentaje Global	60.4%	39.6%	80.2%

Podemos observar que el porcentaje de acierto del grupo de reserva es muy elevado por lo que podemos decir que el modelo es bueno.

La curva ROC muestra que el área bajo la curva es de 0.795. El cual es un valor próximo a uno, por lo que se puede decir que el modelo tiene una buena capacidad de clasificación, como se muestra en la figura 3.

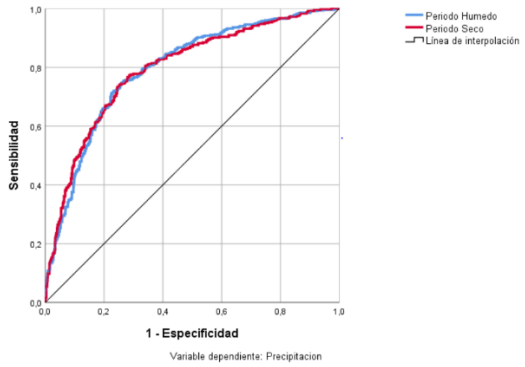

Fig. 3. Curva ROC del modelo

Si ejecutamos una prueba con todos los datos que tenemos disponibles podremos observar cómo se comporta el modelo en una prueba real los resultados que se obtuvieron se muestran en la tabla 12:

TABLA 12
Tabla de clasificación de datos de test con todos los datos

Test	Pertenencia a grupos pronosticada		
Observado	Periodo Húmedo	Periodo Seco	Porcentaje Correcto
Periodo Húmedo	459	112	80.4%
Periodo Seco	135	265	66.3%
Porcentaje Global	61.2%	38.8%	74.6%

Como se puede observar 459 datos fueron clasificados correctamente en el periodo húmedo y 265 en el periodo seco así mismo el porcentaje que obtenemos es menor al que teníamos en la prueba anterior, sin embargo, al comparar la curva ROC y su área 0.801 podemos notar que el modelo establecido funciona, lo cual es reflejado en la figura 4.

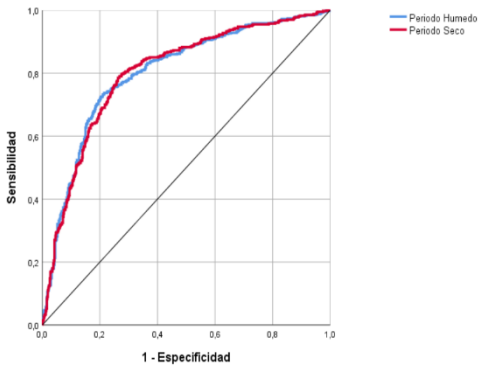

Fig. 4. Curva ROC de la prueba con todos los datos

Revisando los resultados de esto, observamos que los porcentajes de clasificación correcta son altos, lo que indica que el modelo usado con la red es confiable y logro capturar la relación de dependencia presentes en los conjuntos de datos analizados. El porcentaje de clasificación correcta para el periodo húmedo fue del 80.4%, mientras que para el periodo seco fue de 66.3%, los porcentajes de clasificación errónea fueron de 19.6% y 33.7% respectivamente.

3.3 Análisis de resultados de la comparación entre la red neuronal artificial perceptrón multicapa y el análisis discriminante

Los resultados de clasificación obtenidos por los métodos son distintos debido a que para el análisis discriminante se utilizó solo una variable debido a las restricciones de esta, ya que es necesario que se cumpla la significancia de M de Box, sin embargo, si deseamos probar la misma cantidad de variables y compararlas con la RNA el resultado se muestra en la tabla 13.

TABLA 13
Tabla de validación cruzada de análisis discriminante con todos los datos y variables

Validación Cruzada	Pertenencia a grupos pronosticada			
	Precipitación	Periodo Húmedo	Periodo Seco	Total
Recuento	Periodo Húmedo	350	221	571
	Periodo Seco	113	287	400
%	Periodo Húmedo	61.3	38.7	100.0
	Periodo	28.3	71.8	100.0

Validación Cruzada	Pertenencia a grupos pronosticada			
	Precipitación Seco	Periodo Húmedo	Periodo Seco	Total

Como podemos ver los casos clasificados correctamente fueron reducidos desde 95% a 65.8%, pero este valor nos permite observar que el modelo se vuelve inestable debido a que ninguna variable cumple con la prueba de homogeneidad de m de box.

La ventaja que puede tener las RNA sobre el método tradicional son aparentes pues esta puede aprender y mejorar poco a poco siempre y cuando existan más datos para utilizar en el proceso de aprendizaje y validación.

Como pudimos observar en el modelo inicial del análisis discriminante este posee una mejor clasificación correcta que las RNA, pero si lo comparamos usando la misma cantidad de variables y obviando los problemas de significancia del m de box podríamos decir que el RNA es más eficiente debido a que pudo clasificar mejor el Periodo Húmedo y el Periodo Seco.

Si utilizamos el primer modelo de análisis discriminante para realizar la comparación podemos observar que hay una gran distancia entre el 95% y el 74.6% obtenido por la red neuronal. Además, de que los datos que se recogieron para el entrenamiento, validación, y reserva fueron de forma aleatoria con base en 60%, 30% y 10% por lo que no se probó con la misma cantidad de datos cada modelo.

4 Conclusiones

En el caso de la comparación entre el análisis de componentes principales y la red neuronal artificial de tipo hebbiano, el método tradicional si consiguió una reducción de dimensionalidad en la información de calidad de agua estudiada, ya que permitió asociar las variables en grupos que fueran claramente interpretables. Respecto a las redes neuronales artificiales de tipo hebbiano, esta técnica permitió identificar asociaciones con sentido físico/químico entre las variables analizadas, además los grupos de variables encontradas se ubicaron en todos los componentes a excepción de una en la que solo se encontró una variable. El primer componente solo pudo asociarse con la conductividad, mientras que el segundo, tercer y cuarto componente pudo asociarse con 3 variables.

En conclusión, ambos métodos obtuvieron buenos resultados, pero la RNA de tipo hebbiano pudo encontrar más relación entre las variables estudiadas.

En el caso de la comparación entre el análisis discriminante y la red neuronal artificial perceptrón multicapa El método tradicional pudo demostrar mucho mejor la clasificación de datos debido a que se utilizó sobre una variable, la limitación de este método, a pesar de obtener un valor de 95%, es que se necesitan cumplir todos los requisitos para poder ejecutarla si no los datos obtenidos no son estables. En comparación con la red neuronal perceptrón multicapa que es más flexible, también obtiene una clasificación correcta de un porcentaje menor, pero que puede mejorar si se emplean un entrenamiento con un número mayor de patrones, además de que es un sistema tolerante a fallos.

De acuerdo a los resultados obtenidos en este trabajo, los procedimientos de análisis multivariado empleados comúnmente en el análisis de la información de calidad de agua pueden brindar resultados que pueden ser poco interpretables, o muy sensibles a la presencia de valores anómalos. Estos problemas pueden ser resueltos con la utilización de metodologías no lineales como las redes neuronales artificiales de perceptrón multicapa, las cuales permiten construir información que dependen de las relaciones entre el conjunto de datos, y las cuales no ponen restricciones con respecto a la información original. Así mismo, estos modelos son robustos frente a la presencia de valores anómalos.

Cabe mencionar que, en el análisis exploratorio de los datos encontrados en la red de salud Abancay. se observó variables con datos ausentes en su totalidad cuyos fueron, bacterias coliformes totales, bacterias coliformes fecales, bacterias heterotróficas, por lo cual se optó a utilizar solo las variables que contaban con datos en todos los campos las cuales fueron cloro, conductividad, pH, turbiedad, además de la variable periodo que se utilizó de acuerdo a lo mencionado anteriormente en el estudio. Por lo que es recomendable para la red de salud Abancay, conseguir equipamiento que permita recoger más datos que expresen las demás variables físico-químicas.

Es necesario realizar más investigaciones utilizando métodos convencionales y las redes neuronales artificiales para poder determinar que herramienta es más eficaz para asociar y clasificar datos, teniendo en cuenta, que en este último existen una variedad de redes por aplicar.

Referencias

[1] M. Greenacre, P.J.F. Groenen, T. Hastie, et al, "Principal component analysis", Nat Rev Methods Primers vol. 2, pp. 100, 2022, https://doi.org/10.1038/s43586-022-00184-w

[2] S. Mishra, U. Sarkar. et al, "Principal Component Analysis," International Journal of Livestock Research, 2017, https://doi.org/10.5455/ijlr.20170415115235

[3] D. Dhamnetiya, MK. Goel, RP. Jha, S. Shalini, K. Bhattacharyya, "How to Perform Discriminant Analysis in Medical Research? Explained with Illustrations," J Lab Physicians, vol. 14(4), pp. 511-520, 2022 https://doi.org/10.1055/s-0042-1747675

[4] R. Flóres y J. Fernández, "Las Redes Neuronales Artificiales: Fundamentos teóricos y aplicaciones prácticas," La Coruña, NETBIBLO. 2008 S. L, 2008. ISBN: 978-84-9745-246-5.
[5] M. W. Gardner and S. Dorling, "Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences," Atmospheric environment, vol. 32, pp. 2627–2636, 1998
[6] R. Aroca, "Redes neuronales para el análisis y control de calidad del aceite de oliva virgen extra," Universidad Complutense de Madrid, Madrid, 2019.
[7] N. Rodríguez, "Pronóstico de Demanda de Agua potable mediante Redes Neuronales," Universidad Técnica Federico Santamaria, Valparaiso, 2016.
[8] S. Gour y M Gour, "Neural Network Approach In Water Quality Data Analysis For The River Narmada,", Binary Journal of Data Mining & Networking, vol. 4, pp. 49-53, 2014.
[9] S. Heddam, A. Bermad y N. Dechime, "Applications of Radial-Basis Function and Generalized Regression Neural Networks for Modeling of Coagulant Dosage in a Drinking Water-Treatment Plant: Comparative Study," Journal of Environmental Engineering, vol. 137, pp. 1209-1214, New York, 2011, https://doi.org/10.1061/(ASCE)EE.1943-7870.0000435
[10] J.C. Rodríguez, "Estudio Comparativo de Técnicas Estadísticas Multivariadas versus las Redes Neuronales Artificiales en el análisis de datos de Calidad de Agua," Pontificia Universidad Javeriana, Bogotá, 2008.
[11] J. Valencia, "Estudio Estadístico de la Calidad de las aguas en la cuenca hidrográfica del río Ebro," Universidad Politécnica de Madrid, Madrid, 2007.
[12] R. Padilla, "Modelo de Red Neuronal para mejorar la Dosificación de Cloro Gas en la planta de tratamiento de agua potable de la municipalidad provincial de Tayacaja," Universidad Nacional del Centro del Perú, Huancayo, Perú, 2021.
[13] A. Peña, L. Flores del Pino, "Redes neuronales para el tratamiento de agua potable en zona de altitud del Perú," Ambiente y Desarrollo vol. 18, pp. 109-116, 2014, https://doi.org/10.11144/Javeriana.AyD18-35.rnta
[14] R. Álvarez, "Estadistica mutivariante y no parametrica con SPSS. Aplicacion a las ciencias de la salud," Ediciones Diaz de Santos, S. A., Madrid, 1995. ISBN: 978-84-7978-180-4.
[15] M. Andersson, J. Palm, "Forecasting the Stock Market - A Neural Network Approach," MÄLARDALEN UNIVERSITY, Västerås, 2009.
[16] L. Díaz, "Estadística Multivariada: Inferencia y Métodos", Bogota, 2007. ISBN 978-958-701-195-1.
[17] D. Hahs-Vaughn, "Applied Multivariate Statical Concepts," New York, 2017. ISBN: 978-1-315-81668-5.
[18] J. Hilera. y V. Martínez. "Redes Neuronales Artificiales: Fundamentos, modelos y aplicaciones," Madrid, 1995. ISBN: 84-7897-155-6.
[19] P. Isasi y I. Galván, "Redes de Neuronas Artificiales: Un enfoque práctico," Madrid, 2004. ISBN: 84-205-4025-0.
[20] R. Johnson y D. Wichern, "Applied Multivariate Statistical Analysis," vol. 6a, New Jersey, 2007. ISBN 13: 9780131877153.
[21] J. Rodríguez, M. Ferreras y A. Núñez, "Inferencia Estadística, niveles de precisión y diseño muestral," Madrid, 1991.
[22] C. Sierra, "CALIDAD DEL AGUA – Evaluación y diagnóstico," Medellín, 2011. ISBN: 978-958-8692-06-7.
[23] Y. Vidaurre, "Aplicación de las redes neuronales artificiales para el pronóstico de la demanda de agua potable en la empresa EPSEL S.A de la ciudad de Lambayeque," Universidad Católica Santo Toribio de Mogrovejo, Chiclayo, Perú, 2012.
[24] F. Murtaghl, "The Multilayer Perceptron for Discriminant Analysis: Two Examples," Springer, Berlin, Heidelberg, 1992 doi: https://doi.org/10.1007/978-3-642-46757-8_32
[25] L. Heazlewood, J. Walsh, M. Climstein, J. Kettunen, K. Adams, M. DeBeliso, "A Comparison of Classification Accuracy for Gender Using Neural Networks Multilayer Perceptron (MLP), Radial Basis Function (RBF) Procedures Compared to Discriminant Function Analysis and Logistic Regression Based on Nine Sports Psychological Constructs to Measure Motivations to Participate in Masters Sports Competing at the 2009 World Masters Games," Advances in Intelligent Systems and Computing, vol. 392, Springer, Cham, 2016, doi: 10.1007/978-3-319-24560-7_12
[26] E.I. Altman, G. Marco, F. Varetto, "Corporate distress diagnosis: Comparisons using linear discriminant analysis and neural networks (the Italian experience)," Journal of Banking & Finance, vol. 18, pp. 505-529, 1994, https://doi.org/10.1016/0378-4266(94)90007-8
[27] K. Diamantaras, "Robust hebbian learning and noisy principal component analysis," International Journal of Computer Mathematics, vol. 67, pp. 5-24, 1998, https://doi.org/10.1080/00207169808804649

BIOGRAFÍA

Frank Michael Zuloaga Estacio, bachiller en Ingeniería Informática y Sistemas de la Universidad Nacional Micaela Bastidas de Apurímac.

Mario Aquino Cruz, Docente en la Universidad Nacional Micaela Bastidas de Apurímac - Perú, MSc. en Informática, investigador en las áreas de informática educativa, IoT, inteligencia artificial y ciberseguridad.