ISSN: 2710 - 0901 | ISSN-L: 2710 - 0901

https://revistarebi.org/

REBI. Revista Boliviana de Ingeniería
Volumen 6 | No. 2 | Julio - diciembre 2024
http://doi.org/10.62319/rebi.v.6i2.20
Páginas 33 - 46

Diseño de un modelo predictivo mediante redes neuronales para la contaminación ambiental en el Carabayllo-Lima

Design of a predictive model for environmental pollution in Carabayllo-Lima using neural networks

Raul Trinidad Jacinto Herrera
jacintora34@gmail.com
https://orcid.org/0000-0001-7556-6365
Universidad Nacional Federico Villarreal
Lima, Perú.

Recibido el 11 de marzo de 2024 / Arbitrado el 30 de abril de 2024 / Aceptado el 06 de junio de 2024 / Publicado el 01 de julio de 2024

RESUMEN

En la actualidad, Perú enfrenta graves problemas ambientales que impactan directamente la salud de sus habitantes, especialmente en el Distrito Carabayllo-Lima, donde la polución representa una preocupación significativa. Con el propósito de abordar esta problemática, se plantea el uso de redes neuronales para desarrollar indicadores predictivos basados en datos de la Estación de Calidad de Aire local. Este estudio de nivel explicativo y diseño experimental evaluó tres algoritmos de retropropagación y dos modelos de neuronas en una capa oculta durante un período de 2 años. Los resultados obtenidos indican que el modelo propuesto logró una precisión del -0.1089% en la predicción de la calidad del aire, lo que sugiere su utilidad como herramienta para respaldar las decisiones municipales en la gestión ambiental efectiva.

Palabras Claves: Redes Neuronales Artificiales, Predicción Ambiental, Series de tiempo, Contaminación Ambiental; Material Particulado.

ABSTRACT

Currently, Peru is facing severe environmental issues that significantly impact the health of its population, particularly in the Carabayllo-Lima District where pollution is a major concern. To address this challenge, the use of neural networks to develop predictive indicators based on data from the local Air Quality Station is proposed. This explanatory level and experimental design study assessed three backpropagation algorithms and two neuron models in a hidden layer over a 2-year period. The results indicate that the proposed model achieved a precision of -0.1089% in predicting air quality, suggesting its utility as a tool to support municipal decision-making in effective environmental management.

Keywords: Artificial Neural Networks, Environmental Prediction, Time Series, Environmental Pollution; Particulate Matter.

INTRODUCCIÓN

El avance sostenido de la contaminación ambiental a la que estamos expuestos, que se hace más crítico a medida que la población urbana crece, ha conllevado a la investigación y diseño de modelos alternativos para predecir los niveles de los agentes químicos que son contaminantes y que tienen efectos perjudiciales sobre el medio ambiente, ocasionando diversos daños. Los daños son tan diversos que influyen en la economía, adjudicando daños a la propiedad, la biodiversidad, el consumo, al ocasionar enfermedades y más. Esta problemática ha venido también evolucionando progresivamente, haciéndose necesaria estudiar esta situación de una manera particular en zonas críticas para tratar, mediante modelos predictivos, y así dar un manejo de datos y estrategias eficientes para reducir este problema a fin de contribuir en la protección a la salud y al ambiente.

En este sentido, la Organización Mundial de la Salud (OMS) reporta que Lima es la segunda ciudad con mayor contaminación atmosférica en Latinoamérica, con un valor promedio anual de material particulado (PM 2.5) de 48 µg/m3.1. Por ello, es importante generar evidencias destinadas a reducir la magnitud e impacto de la contaminación atmosférica. De acuerdo a la reflexión anterior, se puede decir que la contaminación del ambiente en Perú es bastante significativa en los últimos años, se ha acentuado aún más en la provincia de Lima Metropolitana en el Distrito de Carabayllo, en donde son recurrentes los problemas de salud en sus moradores por temas relacionados a afecciones provenientes del aire por material particulado (MP) debido a diversos factores, como la presencia de muchas calles y avenidas sin asfalto o con capa asfáltica deteriorada, ausencia y/o déficit de áreas verdes en diversas bermas centrales de avenidas.

También por la presencia de canteras de explotación de minerales no metálicos en las zonas periféricas del distrito, lo que produce un tránsito constante de camiones con minerales no metálicos procesados por las principales avenidas del distrito. Además, la topografía del distrito hace que en la falda de los cerros (estribaciones o contrafuertes andinos) se acumule los contaminantes que provienen de la flota vehicular de Lima Metropolitana.

De acuerdo con Sánchez et al. (2008) la São Paulo State Environmental Protection Agency, CETESB (Compañía de Tecnologia de Saneamento Ambiental) afirma que el 97% de todas las emisiones de hidrocarburos son emitidos por vehículos, y el 40% de todas las emisiones de material particulado inhalable (PM10) provienen de fuentes móviles, con base en esto puede estimarse que también el material particulado PM2.5 provenga de fuentes móviles.

Para la zona en mención, la contaminación ambiental originada por material particulado viene a ser un reclamo permanente de la población, es a raíz de esto y en concordancia con sus políticas de desarrollo que el SENAMHI, en trabajo coordinado con la Municipalidad de Carabayllo, instala en el mes de junio de 2014 una Estación automatizada de Calidad de Aire en el distrito de Carabayllo; la cual ha permitido tener una serie de datos continuados sobre ciertos parámetros de calidad de aire como son: i) material particulado PM2.5, ii) material particulado PM10, iii) Dióxido de Azufre, iv) Monóxido de Carbono, v) Ozono y; vi) Dióxido de Nitrógeno.

Para un mejor manejo de las políticas ambientales de los diferentes niveles de gobierno, se hace necesario tener una proyección de lo que pueda ocurrir en materia de problemas ambientales, siendo así relevante establecer métodos de pronóstico de variables de contaminación ambiental para tomar decisiones proactivas ante los problemas de contaminación mencionados.

En este punto debemos señalar que, si bien el presente estudio se hace para evaluar la predicción de variables de contaminación ambiental en la zona de estudio (Carabayllo-Lima-Lima), el enfoque metodológico puede aplicarse para predecir las variables de contaminación ambiental en cualquier estación automatizada de calidad de aire de Lima Metropolitana, así como en cualquier región del país siempre que exista una provisión suficiente de datos que permita hacer uso de la técnica de redes neuronales propuesta.

Para la proyección de datos existen diversas técnicas, entre las cuales las redes neuronales artificiales han demostrado ser capaces de realizar una proyección más eficiente de datos futuros. Al tener datos monitoreados en forma continua, es posible obtener una serie de tiempo en la cual las variables de contaminación ambiental por material particulado PM10 y PM2.5 y contaminantes químicos (SO2, CO, O3 y NO2) varían en función al tiempo transcurrido. El trabajo se realizará considerando las variaciones estacionales que pudieran ocurrir y ofrece como resultado valores futuros cercanos de las variables de contaminación ambiental que permitan establecer políticas preventivas para remediar situaciones no deseadas.

El presente estudio se enfoca a la investigación de una zona geográfica específica, ubicada a la altura del km 16 a 22 de la Carretera Lima-Canta. La investigación pretende demostrar que es posible hacer un pronóstico adecuado para las emisiones de material particulado mediante el uso de la técnica denominada redes neuronales artificiales, bajo la premisa de que es posible hacer una predicción de los parámetros de calidad del aire para la zona de estudio, se propone diseñar y probar un modelo matemático basado en redes neuronales artificiales que sea capaz de predecir con un margen razonable de error, las proyecciones de las variables de contaminación ambiental.

Siendo el objetivo establecer los parámetros del modelo de red neuronal artificial (RNA) del tipo perceptrón multicapa más adecuado que permita realizar este pronóstico y por ende ayudar en forma proactiva a tomar las decisiones más adecuadas para la mitigación de este problema ambiental y con ello los problemas de salud que genera. Para la realización de la meta trazada primeramente se determinaron los parámetros óptimos del modelo de redes neuronales artificiales tipo perceptrón multicapa que permitan obtener indicadores confiables de contaminación ambiental. Posteriormente se obtuvo un pronóstico adecuado de contaminación ambiental por material particulado PM2.5 mediante una red neuronal del tipo perceptrón multicapa con sus parámetros optimizados, y para finalizar se compararon los valores pronosticados de PM2.5 con los Estándares de Calidad Ambiental y emitir las recomendaciones técnicas para que los entes competentes del gobierno local tomen las decisiones adecuadas para mitigar problemas de contaminación ambiental.

MÉTODO

El presente estudio responde a una investigación explicativa, ya que fue basada en la búsqueda del porqué de los hechos mediante el establecimiento de relaciones causa-efecto. En este sentido en la investigación se determinó que la causa es la contaminación del aire en el Perú y por ende en el distrito de Carabayllo y en el caso de los efectos corresponde a la demostración de la efectividad del modelo basado en redes neuronales artificiales capaz de pronosticar variables de contaminación ambiental de material particulado con buena precisión y en una forma sencilla.

El estudio se realizó bajo un diseño experimental puro ya que se controlaron todos los factores que pudieran alterar el proceso, en el modelo se emplearon grupos de comparación y sus equivalencias

mediante la asignación aleatoria, además pasó un proceso de validez interna, es decir, que se garantizó que los resultados son producto de la variable independiente y no de otros factores y por último pasaron por una validez externa en las que se deja abierta la posibilidad de generalizar los resultados a otras localidades y en otros ambientes.

Variable Independiente (X): Modelo de Predicción basado en Red Neuronal Artificial tipo Perceptrón Multicapa.

Indicadores: Están basados en fenómenos físicos (Número de variables de entrada (1 a N) y número de variables de salida (1 a N)). Basados en el modelo matemático (Número de capas (1 a N), número de retrasos (1 a N), Tipo de funciones de activación (sigmoidal, tangente hiperbólica, etc.), pesos (Reales mayores que cero), tipo de predicción (etapa única o multietapa)).

Variable Dependiente (Y): Resultados obtenidos mediante la aplicación del modelo de predicción (valores futuros cercanos de concentración de Material Particulado PM2.5 en ppm).

Indicadores: Error de Correlación, Error Mínimo Cuadrático (MSE o Mean Square Error, por sus siglas en inglés) y Error Porcentual Medio (MPE o Mean Percentage Error, por sus siglas en inglés).

La población de estudio estuvo constituida por las diez (10) Estaciones de Calidad del Aire instaladas por el Servicio Nacional de Meteorología, Hidrología y Navegación (SENAMHI) en la ciudad de Lima Metropolitana. Se ha incluido en esta población solo a las estaciones que generan información que permite gestionar la calidad de aire y generan un flujo continuo de datos de manera automatizada estando ubicadas en la provincia de Lima.

Para esta investigación se ha considerado a una de las estaciones de calidad de aire que cumplen las condiciones de estar dentro de la provincia de Lima y son de la clase automatizada: Estación de Calidad de Aire de Carabayllo, instalada el año 2014 por el SENAMHI en el distrito de Carabayllo mediante un convenio de colaboración con el gobierno local del distrito.

Instrumentos de Recolección de Datos

Dentro de Lima Metropolitana el SENAMHI posee una red de monitoreo de calidad de aire con estaciones dotadas de equipos automáticos que monitorean de manera constante los contaminantes denominados: material particulado, PM10; material particulado, PM2.5; dióxido de azufre, SO2; dióxido de nitrógeno, NO2; ozono, O3 y monóxido de carbono, CO. Entre las estaciones automatizadas establecidas en Lima Metropolitana se tiene como muestra a la ubicada en el distrito de Carabayllo, de la cual se han obtenido los datos utilizados en el presente proyecto de investigación. En la estación de Carabayllo se han instalado los siguientes equipos: 1). Analizadores de contaminantes químicos (Analizador de Ozono (O3), Analizador de Óxido de Nitrógeno (NO), Analizador de Dióxido de Azufre (SO2), Analizador de Monóxido de Carbono (CO)). 2) Analizadores de contaminantes particulados (Analizador de PM10, Analizador de PM2.5).

Para la recolección de datos referidos a partículas se cuenta con la tecnología automatizada basada en monitoreo continuo de partículas en el ambiente. Los equipos de análisis automático continuo toman mediciones continuas directas de masa de las partículas mediante el uso de un instrumento denominado micro balanza de oscilación de elementos cónicos (tapered element oscillating microbalance, TEOM por sus siglas en inglés). Este tipo de monitores se usan para medir partículas aerotransportadas con excelente precisión de corto término.

Mecanismo de funcionamiento: Microbalanza oscilatoria de elementos cónicos, la muestra de aire pasa a través de un filtro que es parte de un sistema que vibra a su resonancia característica. El material particulado colectado sobre el filtro aumenta la masa vibrante y por lo tanto decrece la frecuencia de oscilación en forma proporcional. La concentración del material particulado es calculado a partir de una calibración que relaciona la frecuencia de vibración y la cantidad de material particulado, teniendo en cuenta el volumen de muestra de aire.

Procesamiento y Análisis de Datos

Considerando el marco teórico, se eligió la herramienta de inteligencia artificial denominada red neuronal artificial tipo perceptrón multicapa bajo el modelo no lineal autoregresivo con entrada exógena. Esto debido a su capacidad de predecir el comportamiento de las series de tiempo aplicadas al caso de variables de contaminación ambiental del aire clasificadas como material particulado.

Procesamiento inicial de datos de entrada

Importación de Datos: Para el trabajo inicial con el modelo NARX (no lineal autoregresivo con entrada externa) es necesario importar cinco columnas de datos. Mediante estos comandos se obtiene un vector de 2,888 filas por 5 columnas, este incluye elementos vacíos producto de omisiones de registro en la estación automatizada, por lo que es necesario eliminar los registros sin valores. Para esto deben trasladarse los datos numéricos recibidos en formato de hoja de cálculo Excel hacia el formato especial de datos utilizado por Matlab. El siguiente ejemplo de código Matlab permite este proceso para la variable PM10 y se repite un procedimiento semejante para cada uno de los contaminantes.

Eliminación de registros sin valores

Mediante los siguientes comandos Matlab eliminamos los registros que contienen un campo sin valor (NaN, not a number). Esto permitirá asegurar un procesamiento de datos sin incurrir en distorsiones que se producen al procesar campos vacíos. 1) % H es un vector que contiene los valores importados, 2) % Eliminación de Filas que tengan NaN en cada columna.

La aplicación del comando isnan permite la eliminación de los datos NaN, para realizar el procesamiento de datos solamente sobre valores positivos, eliminando los campos que contienen los datos NaN. Se obtiene de esta manera un vector de 820 filas por 5 columnas de datos válidos para el procesamiento posterior. Es decir, no hubo más datos eliminados, y se trabajará con un vector de valores positivos de PM2.5.

Normalización de las entradas (Transformación Lineal)

Se debe considerar que las variables de entrada pueden tener diferencias de valores de varios órdenes de magnitud de forma que el aprendizaje de la red se verá influenciado por estas diferencias, ya que el incremento de pesos de una neurona es proporcional a su entrada. Este problema puede evitarse asignando valores parecidos a las variables de entrada mediante la normalización de sus valores. Para el presente proyecto, se plantea la utilización de la normalización que transforma los datos al rango [-1 +1] mediante el siguiente segmento de programa Matlab que realiza la transformación lineal para PM10 y semejantes para otros contaminantes.

Patrones de entrada (input de la red)

Los factores que influyen directamente en la predicción de material particulado PM2.5 son: valores anteriores de PM2.5, valores de otros contaminantes químicos, porcentaje de humedad relativa, temperatura, contaminantes químicos, entre otros. Debe considerarse que para el caso de estudio aún no se cuenta con información de datos meteorológicos. Desde que se ha observado que uno de los valores que influyen fuertemente son los propios valores anteriores del material particulado PM2.5, así como los de otros contaminantes por material particulado PM10 y contaminantes químicos; y debido a la naturaleza del proyecto de investigación se considera solo como patrones de entrada para la red neuronal artificial tipo perceptrón multicapa, a los valores anteriores de material particulado PM2.5 y de otros contaminantes químicos.

Estos elementos de los patrones de entrada se han ordenado como un arreglo de vectores (inicialmente cinco columnas por 820 filas) para alimentar a la red y utilizar las capacidades de aproximador universal de funciones de estas para estimar la función de salida.

Patrones de salida (output de la red)

Con base en los patrones y las muestras de datos históricos de la concentración de material particulado PM2.5, se procedió a evaluar la influencia de los patrones de entrada sobre el comportamiento de la concentración de material particulado PM2.5. Con esto se obtiene un arreglo de vectores de dimensión 1xN como patrón de salida, lo cual representan los valores estimados de material particulado PM2.5.

Determinación del número de capas ocultas de la red

No existe una metodología específica para determinar el número de capas ocultas (hidden layer) en una red neuronal artificial. Sin embargo, cuando el número de capas es mayor a uno, la capacidad de discriminación en el espacio RN se hace más potente, aun cuando en el caso de N=3, esta capacidad no necesariamente aumenta con más capas ocultas. De este hecho se puede deducir que las capas ocultas variarían entre un mínimo de una y un máximo de tres capas ocultas. La figura 20 presenta los espacios de discriminación que se pueden lograr con diversos números de capas ocultas.

Para el presente caso se formuló la arquitectura inicial con dos capas ocultas y con un número de neuronas en relación de 2:1 entre la primera y segunda capa. Sin embargo, debido a los extensos tiempos de entrenamiento se decidió por utilizar una sola capa oculta donde las neuronas tuvieran la forma de vectores [1x1] y [1x2].

Determinación del número de neuronas por capa

Para iniciar el cálculo se utilizó un múltiplo de 2 en la única capa oculta, se varió el tamaño de los vectores utilizados en las neuronas de la capa oculta utilizando [1x1] y [1x2]. Se han introducido retrasos (delays) en el vector de entrada y en el de realimentación para los valores de concentración de PM2.5 estimada, desde que el valor actual de concentración está directamente relacionado con los N últimos valores registrados.

Para determinar el número óptimo de neuronas en las capas ocultas se ha realizado un proceso de ensayo-error, siendo el objetivo diseñar el número adecuado de neuronas en ambas capas ocultas para que estas pudieran aprender las características de las potenciales asociaciones (relaciones) entre los datos muestrales.

Las funciones de activación (funciones de transferencia) utilizadas para la red neuronal artificial fueron: tansig y purelin. Los retrasos en el caso del modelo propuesto de red neuronal artificial hacen que los datos de ingreso no solo dependan de variables externas, sino también de sí mismo en periodos anteriores. La colocación de retrasos a los valores de concentración de PM2.5, tanto a la concentración estimada como a los de los periodos anteriores y el valor actual de concentración de PM2.5, están directamente relacionado con los últimos N valores registrados.

Inicialmente se tomaron dos retardos para la formulación de la red neuronal, aun cuando este valor varió de acuerdo a los ensayos y simulaciones realizadas, utilizando también cuatro retardos, tanto en la capa de input como en la de feedback.

Funciones de activación

Desde que se conoce que las funciones de activación tienen usos determinados, tal como se puede apreciar:

- ✓ Sigmoide y tangente hiperbólica: redes de predicción y redes de pronóstico de procesos que modelan sistemas dependientes del tiempo.
- ✓ Función de base radial: redes de clasificación.

Luego, y como estamos en un caso de predicción seleccionamos la función de activación más adecuada como es la tangente hiperbólica (tansig, en Matlab). Aplicamos esta función para la única oculta. Para la función de activación de salida en cambio se utilizó la función lineal (purelin, en Matlab).

Una vez definido el tipo de red neuronal artificial a ser utilizado (red neuronal artificial tipo perceptrón multicapa con retrasos), así como sus parámetros y conjuntos de datos de entrada y de salida, se procedió a realizar el diseño de la arquitectura de la RNA. Considerando las conexiones de este tipo, los retrasos y la retroalimentación convierten a esta red en un modelo dinámico.

Con base en los conceptos enunciados y utilizando el programa fuente Matlab versión R2015 ha diseñado para la predicción de multietapa adelantada.

Técnicas y Procedimientos de Análisis de Datos

En una fase inicial se formularon las herramientas matemáticas de inteligencia artificial denominada redes neuronales artificiales tipo perceptrón multicapa que permitió ejecutar el modelo predictivo de multietapa adelantada deseada con el mínimo error posible. Para probar la hipótesis se utilizó la herramienta estadística denominada Error Cuadrático Medio, posteriormente se midió el *Error de Pronóstico* con la finalidad de conocer que tan adecuado puede ser un modelo o una técnica para pronosticar; la decisión para utilizar una técnica de pronóstico en particular, depende de si la técnica producirá errores de predicción que se valoren como suficientemente pequeños. Los siguientes indicadores evalúan el nivel de precisión de cada método o técnica: Error Porcentual Absoluto Medio (MAPE) y Error Porcentual Medio (MPE).

Una vez seleccionada la muestra y a través de la Gerencia de Servicios a la Ciudad y Medio Ambiente, se solicitó al SENAMHI los datos de calidad del aire referidos a los años 2017 y 2018, los cuales fueron proporcionados en formato de hoja de cálculo.

Dadas las características especiales que reviste el procesamiento de datos con la técnica de redes neuronales artificiales, se utilizarán la totalidad de los datos horarios recabados (es decir se utilizará el íntegro de los datos), exceptuando los registros que poseen campos vacíos.

Son 2.888 registros equivalentes a información de cada hora con datos de los contaminantes que han sido registrados en la estación de calidad de aire de Carabayllo, los contaminantes fueron los siguientes: Material particulado, PM10, Material particulado, PM2.5, Dióxido de azufre, SO2, Dióxido de nitrógeno, NO2, Ozono, O3 y Monóxido de carbono, CO.

Cabe destacar que, a lo largo de la operación de la estación de calidad de aire, en varias ocasiones por diversos motivos se detuvo el registro de los datos, por ende, existe una cantidad del total de registros, tanto en los datos de material particulado PM2.5 como en otros datos, que están sin valor. Para estos casos se utilizará el software Matlab R2015a para eliminar estos campos vacíos, esto con el objeto de que cada fila de datos considerada tenga un valor en cada uno de los contaminantes, de modo que se pueda hacer el procesamiento posterior sin la dificultad que representa la manipulación de datos vacíos en alguna columna. Después de este procesamiento se obtuvo una hoja de cálculo con datos distribuidos en cinco columnas por 820 filas que involucran datos completos para cada celda.

Procedimientos para la ejecución del estudio

El desarrollo del presente estudio consta de una secuencia de cuatro etapas, la primera constó del levantamiento y preparación de los datos; los relacionados de campo se obtuvieron en formato de hoja de cálculo el cual fue generado en la Estación de Calidad de Aire de Carabayllo y su pre-procesamiento se realizó mediante el software Matlab.

La segunda fase se trató de la recopilación de la información que incluyó la recolección de libros, tesis, artículos tanto en medio físico como virtual; así como consulta a páginas web de instituciones vinculadas al tema en estudio. Esta información se analizó, se clasificó, se catalogó y se formó el respaldo bibliográfico para la tesis en curso.

La tercera fase fue la construcción del modelo predictivo de multietapa adelantada (multistep ahead), basada en las redes neuronales artificiales denominadas perceptrón multicapa con la técnica denominada no lineal auto regresiva con entrada exógena (NARX) como propuesta de trabajo.

Y la cuarta y última parte fue la validación del modelo de predicción, mediante comparación del modelo predictivo con datos históricos reales en la estación de estudio y estimación del margen de error.

RESULTADOS

Se diseñaron y desarrollaron tres diferentes modelos de redes neuronales artificiales del tipo perceptrón multicapa en la versión R2015a de Matlab, mediante la utilización de los siguientes parámetros de entrada: material particulado PM_{10}, material particulado $PM_{2.5}$, dióxido de azufre SO_2, dióxido de nitrógeno NO_2, monóxido de carbono CO; y un solo parámetro de salida: material particulado $PM_{2.5.}$ Los modelos de redes neuronales artificiales del tipo perceptrón multicapa se entrenaron utilizando los siguientes tres algoritmos de retropropagación: 1. Levenberg-Marquardt (LM), 2. Regulación Bayesiana (BR) y Scaled Conjugate Gradient (SCG). Asimismo, se entrenaron siguiendo dos estructuras de datos para las neuronas de la capa oculta: los vectores tipo [1x1] y [1x2].

Se diseñaron los siguientes seis grupos de experimentos, cada uno de los cuales se incluyó un barrido de ocho simulaciones para obtener el menor valor del criterio de evaluación seleccionado. Estas evaluaciones fueron ejecutadas mediante un programa en Matlab versión R2015a y el criterio de evaluación seleccionado está dado por el Error Cuadrático Medio (MSE):

Cuadro 1

Diseño de las simulaciones para obtener parámetros óptimos de red

Algoritmo de Entrenamiento	Capas Ocultas	Neuronas [1x1]	Neuronas [1x2]
Levenberg-Marquardt	1	2 hasta 16	[2 1] hasta [16 8]
Regulación Bayesiana	1	2 hasta 16	[2 1] hasta [16 8]
Scaled Conjugate Gradient	1	2 hasta 16	[2 1] hasta [16 8]

Con base en los valores de error medio cuadrático (MSE) (función de que mide el rendimiento de la red de acuerdo a la media de los errores cuadráticos. En Matlab se calcula mediante: Perf = mse [red_neural, target, output]) de las simulaciones basadas en la red neuronal artificial tipo perceptrón multicapa con variación del algoritmo de entrenamiento expresado en las tablas anteriores, seleccionamos los mínimos de estos valores y se presenta el resumen de estos en la tabla siguiente.

Cuadro 2

Menores valores de Error Medio Cuadrático

MENORES VALORES DE ERROR MEDIO CUADRATICO (MSE) PARA LOS DIFERENTES ESCENARIOS			
ALGORITMOS DE ENTRENAMIENTO / CAPAS OCULTAS	LEVENBERG-MARQUARDT	REGULACION BAYESIANA	SCALE CONJUGATED GRADIENT
UNA CAPA OCULTA [1x1]	0.001280	0.002512	0.001009
UNA CAPA OCULTA [1x2]	0.001924	0.002837	0.001693
MSE MÍNIMO		0.001009	

En virtud de los resultados anteriores del valor mínimo de error medio cuadrático (MSE mínimo) se identificaron los parámetros del modelo óptimo con los resultados que se presentan a continuación:

Cuadro 3

Parámetros del Modelo Óptimo de RNA

REBI. Revista Boliviana de Ingeniería
Volumen 6, Número 2 / Julio - diciembre 2024 / ISSN: 2710 - 0901 | ISSN-L: 2710 - 0901

PARÁMETROS DEL MODELO ÓPTIMO		
ALGORITMO DE ENTRENAMIENTO		GRADIENTE DE ESCALA CONJUGADA
RETRASOS	INPUT	4
	FEEDBACK	4
CAPAS OCULTAS		1
TIPO DE NEURONAS		[1x1]
NEURONAS EN CAPA OCULTA		8
MSE MÍNIMO		0.001009

Una vez obtenidos los parámetros óptimos del modelo de redes neuronales del tipo perceptrón multicapa para la predicción de la evolución de la concentración de material particulado PM2.5, se utilizaron los valores de las concentraciones de material particulado PM2.5 en el intervalo de 72 horas del futuro cercano. Después de obtener estos resultados y previa transformación lineal inversa, se han comparado estos valores con los estándares de calidad ambiental (ECA) para material particulado PM2.5. En base a este procedimiento se puede enunciar lo siguiente:

- ✓ Con un buen grado de aproximación (error cuadrático medio o MSE de 0.001009), ha sido posible pronosticar valores futuros cercanos que abarcan datos referidos a 72 horas o 3 días, los cuales junto con la transformación lineal inversa y la posterior comparación con las normas de calidad de aire del estado peruano pueden ser utilizadas para diseñar políticas de gobierno local del distrito de Carabayllo que permitan mejorar los episodios de calidad de aire que superen los máximos valores establecidos en los estándares de calidad del aire.
- ✓ Los experimentos fueron llevados a cabo mediante el diseño de una red neuronal artificial del tipo perceptrón multicapa con una sola capa oculta. Se analizaron tres técnicas de entrenamiento mediante retro propagación, realizando dos análisis para cada algoritmo que consistieron en variar el número de neuronas en la capa oculta utilizando vectores de dimensiones [1x1] y [1x2].
- ✓ Mediante corridas previas de la simulación de la red neuronal artificial del tipo perceptrón multicapa y mediante el modelo no lineal autoregresivo con entrada externa, se determinó que el número de neuronas en la única capa oculta varía desde 2 hasta 16 en el caso del vector [1x1] y desde [2 1] hasta [16 8] para el caso de los vectores [1x2].
- ✓ Para los rangos establecidos de simulación se obtuvieron tiempos razonables de entrenamiento, además los mejores resultados siempre se obtienen para casos de un número neuronas del tipo múltiplo de 2 y en especial del formato 2N en la única capa oculta del experimento.
- ✓ Para el experimento se han considerado determinados parámetros de entrada entre los cuales se incluye en forma recursiva al parámetro material particulado PM2.5. La inclusión recursiva del parámetro a pronosticar implica el uso del modelo no lineal autoregresivo con entrada externa.

DISCUSIÓN

Uno de los hallazgos medulares de la presente investigación fue la verificación de que los modelos matemáticos basados en redes neuronales artificiales pueden utilizarse como herramientas de predicción para concentraciones de material particulado PM2.5 en áreas urbanas. Un modelo estadístico basado en redes neuronales artificiales, entrenado con datos específicos en tiempo y ubicación geográfica, puede utilizarse con menos recursos que el uso de un modelo determinístico.

Asimismo, lo pueden constatar los investigadores Salini y Pérez (2006) en su trabajo denominado “Estudio de series temporales de contaminación ambiental mediante técnicas de redes

neuronales artificiales" diseñaron una red neuronal artificial (RNA) para predecir valores de concentraciones horarias de material particulado fino en la atmósfera. Una vez fijo el número de variables de entrada y elegida la variable a pronosticar, se diseñó un modelo predictivo basado en redes neuronales artificiales y los resultados fueron más precisos que los obtenidos con un modelo de persistencia.

Otro punto importante que se pudo determinar en el estudio fue que las redes neuronales artificiales están restringidas a un periodo de tiempo y una determinada ubicación, porque estas siempre requieren ser entrenadas con datos locales. En principio los modelos de redes neuronales artificiales no son adecuados para predecir distribuciones de concentración espacial en grandes áreas, no obstante, el uso de las redes neuronales en forma simultánea en diversas zonas geográficas puede ser útil para establecer un mapa de probables eventos de contaminación en un área geográfica extensa.

Por otra parte, mediante el pronóstico de las variables de contaminación del aire por material particulado PM2.5 se pueden evaluar los resultados de proyectos de mejora de la calidad ambiental, tal como lo expresa la Organización Mundial de la Salud (OMS) al considera a la contaminación del aire como uno de los factores más importantes de riesgo medioambiental para la salud y el bienestar debido a que la exposición a micro-contaminantes en el aire contribuye al desarrollo de diversas enfermedades cerebro-cardiovasculares, respiratorias, oncológicas, entre otras (TFM, 2020).

Asimismo, según el Informe sobre la Calidad del Aire Mundial 2019 elaborado por IQAir el Perú se encuentra en el puesto 33 del ranking mundial de los países más contaminados o con mayor concentración promedio de material particulado. Con una concentración promedio ponderada por habitantes de PM2.5 ascendente a 23 μ/m3, el país se sitúa como el primero de Latinoamérica y el Caribe, por encima de Chile, Guatemala y México. Lima se sitúa en el puesto 15 de las ciudades más contaminadas de Sudamérica, con una concentración promedio anual de PM2.5 ascendente a 23.7 μ/m3 (TFM, 2020), por lo tanto, los resultados de este modelo pueden promover las capacidades de los gobiernos locales para identificar las metas y objetivos concernientes a establecer políticas y diseñar proyectos de mitigación ambiental de la calidad del aire.

Y para finalizar, en la investigación se pudo conocer que la formación de las partículas atmosféricas, tanto primarias (origen antropogénico: procesamiento de minerales, combustión, incendios) como secundarias (reacciones químicas, condensación o coagulación) están influenciadas por las concentraciones de otros contaminantes atmosféricos y ciertas condiciones meteorológicas como porcentaje de humedad y radiación solares. Debe destacarse que en este caso no se tuvo disponibilidad de datos meteorológicos, sin embargo, queda la propuesta de realizar este procesamiento numérico utilizando como datos de entrada al modelo a los parámetros meteorológicos.

CONCLUSIONES

Para ultimar la presente investigación, se planteó la construcción de un modelo matemático, basado en redes neuronales artificiales, para pronosticar valores de contaminación ambiental por material particulado PM2.5 en el distrito de Carabayllo; esta propuesta demuestra un modelo innovador de pronóstico basado solamente en datos, los cuales pueden ser obtenidos en grandes cantidades desde las estaciones automatizadas de calidad de aire, de esta manera se deja por sentado con evidencias

científicas que el modelo puede realizar predicciones en el futuro cercano con un nivel aceptable de desempeño.

En el mismo orden de ideas, los resultados obtenidos demuestran que los modelos basados en redes neuronales artificiales tipo perceptrón multicapa pueden ser aplicado al pronóstico de series de tiempo de calidad ambiental de calidad del aire debido a su capacidad de aprendizaje del tipo supervisado, la única condición es que se debe definir con claridad los parámetros óptimos que permitan obtener resultados adecuados en el pronóstico.

En cuanto a las simulaciones realizadas para los experimentos propuestos se observó que los resultados más precisos se obtuvieron mediante los modelos de redes neuronales del tipo perceptrón multicapa con una sola capa oculta conteniendo un número de neuronas múltiplo de 2 y en especial en los números del tipo 2N.

En el presente estudio se han considerado 6 modelos de RNA y se han evaluado los errores medio cuadráticos para el rango de neuronas en la capa oculta considerada, en general los modelos que tuvieron las neuronas tipo vector [1x1] tuvieron un mejor rendimiento que los del tipo [1x2], todo esto en una única capa oculta. Se destaca de esto que los pronósticos para predicción de calidad del aire pueden ser un referente para tomar decisiones adecuadas y mejorar la calidad ambiental en la zona de influencia.

Con respecto a los diseños, los mejores resultados se obtuvieron con los parámetros del modelo 5 (algoritmo de entrenamiento Gradiente de Escala Conjugado, con 8 neuronas del tipo vector [1x1] en la única capa oculta), seguido de los modelos 1 y 6 respectivamente. Estos modelos son los que poseen los menores errores cuadráticos medios.

En general todos los modelos planteados poseen un nivel de error bajo, lo cual lleva a concluir que los modelos basados en redes neuronales tipo perceptrón multicapa consiguen un buen nivel de aproximación para el pronóstico de valores de contaminación ambiental por materia particulado.

En el caso del modelo 5 con MSE más bajo (MSE=0.001009) presenta además un error porcentual medio bastante bajo (MPE=-0.1089%) lo cual señala que la técnica no presenta mucho sesgo, desde que el valor porcentual MPE es cercano a cero.

De acuerdo a la implementación de la propuesta, el modelo RNA ha sido aplicado al distrito de Carabayllo, mediante los datos de la estación de calidad del aire del mencionado distrito. En principio, su aplicación y uso debe limitarse a esta zona geográfica, no obstante, la metodología propuesta puede utilizarse en otros distritos que tengan estación de calidad del aire, previo entrenamiento de la red para su uso adecuado.

Por otra parte, para garantizar una buena ejecución se deben obtener los parámetros óptimos para los modelos predictivos RNA de contaminación ambiental por material particulado, debido a que para cada área geográfica evaluada las condiciones topográficas y meteorológicas son variables.

En el caso de una ciudad como Lima Metropolitana, por su extensión, se tiene que el área de estudio circunscrita al distrito de Carabayllo configura un microclima tal que se puede considerar un área geográfica aislada para realizar en el presente estudio.

Y para culminar se pudo determinar en la ejecución de la investigación que para conseguir una mejor precisión en las estimaciones de calidad de aire se deben considerarse otros factores, tales como datos meteorológicos, estacionales y la posibilidad de tener datos para un mayor rango de años.

Vale mencionar que en el estudio se han utilizado los datos de dos años completos, lo cual es un conjunto de datos lo suficientemente grande para obtener resultados con cierta confiabilidad, por lo tanto, se estima que existe oportunidad de obtener buenos resultados utilizando un conjunto de datos de menor dimensión, de al menos un año para obtener resultados razonables.

Igualmente, en virtud de los resultados obtenidos se recomienda utilizar la metodología para generar modelos propios en otras ciudades del país donde existan estaciones de calidad del aire. Así se podrá hacer un análisis local de las condiciones químicas que determinan los niveles de concentración de contaminantes atmosféricos en las diferentes zonas del país.

Adicionalmente se debe considerar que existen diferentes técnicas de inteligencia artificial que permiten el tratamiento de datos cuantitativos, por lo que resulta atractivo rediseñar los modelos planteados bajo diferentes técnicas de aprendizaje supervisado, tales como funciones de base radial y algunas variaciones del perceptrón multicapa. Además, también se sugiere el uso de otras técnicas de entrenamiento que van más allá de la retropropagación.

Si bien en el presente trabajo se ha realizado un pronóstico bastante aceptable de las variables de contaminación del aire por material particulado y debido a la simplicidad de su aplicación, se recomienda evaluar su aplicación en diversos sectores de las investigaciones ambientales y verificar si es factible su aplicación.

Finalmente, se asevera que los modelos predictivos de contaminantes de calidad de aire pueden ser optimizados mediante utilización de información meteorológica y topográfica en la zona a estudiar, si se superan las limitaciones de información sobre estas variables analizadas, se recomienda evaluar algunos modelos de redes neuronales artificiales que incorporen estas variables como datos de entrada.

REFERENCIAS

Gonzales, G., Zevallos, A., Gonzales, C., Nuñez, D., Gastanaga, C. y Cabezas, C., (2014). Environmental pollution, climate variability and climate change: a review of health impacts on the Peruvian population. *Revista peruana de medicina experimental y salud pública*; 31(3): 547 - 556. https://www.scielo.org.mx/scielo.php?script=sci_nlinks&pid=S0036-36342017000500050700003&lng=en

Salini, G. y Pérez, P., (2006). Estudio de series temporales de contaminación ambiental mediante técnicas de Redes Neuronales Artificiales. *Ingeniare. Revista chilena de ingeniería,* 284-290. https://www.scielo.cl/scielo.php?pid=S0718-33052006000200012&script=sci_arttext&tlng=pt

Sánchez, O., Ynoue, R., y Droprinchinski, L., (2008). Vehicular particulate matter emissions in road tunnels in Sao Paulo, Brazil. *Envvironmental Monit. Assess. https://link.springer.com/article/10.1007/s10661-008-0198-5*

TFM, (2020). ¿Qué calidad debe tener el aire que respiramos? El impacto del aire en la salud y las condiciones actuales de calidad de aire. https://www.tfm.pe/noticias/la-calidad-del-aire-que-respiramos#:~:text=De%20acuerdo%20al%20informe%20de,ponderada%20por%20habitantes%20de%20PM2.

World Health Organization. (2016). *World Health Statistics 2016 [OP]: Monitoring Health for the Sustainable Development Goals (SDGs)*. World Health Organization. https://books.google.es/books?hl=es&lr=&id=-A4LDgAAQBAJ&oi=fnd&pg=PP1&dq=World+Health+Organization+(WHO)+(2016).+Public+health,+environmental+and+social+&ots=ddjk1VehzE&sig=gu1hJcSngOhX2w9B56SXNNOHSPw