ISSN 1870-4069

Predicción de la tasa de degradación de un proceso de fotocatálisis heterogénea: Comparativa entre una red neuronal artificial y un modelo de bosques aleatorios

Katia L. Rocha-Aguilar, José I. Hernández-Vega,
Cinthia G. Aba-Guevara, Alejandro Alonzo-García,
René Sanjuan-Galindo

Instituto Tecnológico de Nuevo León,
División de Estudios de Posgrado e Investigación,
México

{mg15480487, jose.hv, cinthia.ag, alejandro.ag,
rene.sg}@nuevoleon.tecnm.mx

Resumen. Dada la creciente demanda del consumo de agua, la escasez de este recurso, y el aumento de las aguas residuales generadas por la actividad humana, se han realizado esfuerzos para optimizar, mejorar y proponer tratamientos de aguas residuales alternativos, entre los cuales se encuentra la fotocatálisis heterogénea, un proceso de oxidación avanzado. Varios estudios se han llevado a cabo en un esfuerzo por lograr su implementación a gran escala, sin embargo, el análisis de la eficiencia del proceso es complejo, ya que en el proceso de tratamiento de aguas residuales intervienen diversos factores que resultan complejos de modelar, predecir y/o controlar, y se requiere la realización de un gran número de experimentos. Este artículo propone la implementación de modelos de inteligencia artificial para la predicción de la tasa de degradación fotocatalítica (información relevante para el estudio de la eficiencia del proceso) haciendo uso de TiO2 como catalizador, realizando una comparativa entre un modelo de redes neuronales artificiales y uno de bosques aleatorios.

Palabras clave: Inteligencia artificial, fotocatálisis heterogénea, aprendizaje automático.

Prediction of the Degradation Rate of a Heterogeneous Photocatalysis Process: Comparison between an Artificial Neural Network and a Random Forest Model

Abstract. Given the growing demand for water consumption, the scarcity of this resource, and the increase in wastewater generated by human activity, efforts have been made to optimize, improve, and propose alternative wastewater treatments, including heterogeneous photocatalysis, an advanced oxidation

process. Several studies have been conducted in an effort to achieve its large-scale implementation; however, analyzing the efficiency of the process is complex, since the wastewater treatment process involves various factors that are difficult to model, predict, and/or control, and requires a large number of experiments. This article proposes the implementation of artificial intelligence models to predict the photocatalytic degradation rate (relevant information for studying the efficiency of the process) using TiO2 as a catalyst, comparing an artificial neural network model with a random forest model.

Keywords: Artificial intelligence, heterogeneous photocatalysis, machine learning.

1. Introducción

El aumento del consumo del recurso hídrico para las diferentes actividades humanas representa una mayor generación de aguas residuales, por lo que el impacto ambiental no solo se ve reflejado en la sobre-extracción de los mantos acuíferos, sino también en la contaminación de los ecosistemas debido a los contaminantes vertidos en las corrientes naturales [1]. Es por lo anterior que existen procesos de tratamiento de estas aguas contaminadas, con el fin de reutilizarlas y reducir el impacto ambiental. Representan un riesgo para el medioambiente y la salud la presencia de contaminantes orgánicos en aguas residuales, muchos de los cuales son difíciles de remover o eliminar mediante métodos convencionales de tratamiento o procesos de degradación natural. La degradación por fotocatálisis es una alternativa que ha mostrado poseer un gran potencial en la degradación de estos compuestos, llegando incluso a lograr su completa mineralización [2]. El dióxido de titanio (TiO_2) como fotocatalizador, es uno de los óxidos metálicos ampliamente estudiados por su actividad fotocatalítica y sus propiedades, además de tratarse de un compuesto estable y de barata adquisición [3], sin embargo, su implementación en procesos de fotocatálisis heterogénea presenta varios retos para su implementación a gran escala, entre los cuales se encuentran el estudio de la cinética de degradación y la aplicabilidad para la degradación de diversos contaminantes [4].

El conocimiento sobre las velocidades de reacción bajo determinadas condiciones es esencial para la optimización del proceso, las cuales son analizadas mediante numerosos ensayos experimentales, sin embargo, parámetros que pueden ser óptimos para un determinado tipo de contaminante no son los mismos para otro. Son diversos los factores que impactan en el desempeño de este proceso, como lo son la concentración y tipo de contaminante, la concentración del catalizador, el pH, la fuente de luz, entre otros. Los modelos de inteligencia artificial (IA) se han implementado en diferentes ámbitos de los procesos de tratamientos de aguas residuales, para abordar tareas de optimización, predicción, control y modelado de aguas residuales [5]. Como una subcategoría de estos modelos, se han desarrollado algoritmos de aprendizaje automático con la capacidad de encontrar patrones presentes en los datos y adaptarse al comportamiento no lineal, por ejemplo, de la degradación de contaminantes, realizando esta predicción con base en los parámetros que intervienen en el proceso. En la

Tabla 1. Consideraciones para la selección de los datos.

Tipo de reactor	Reactor Batch	Debido a que las variaciones físicas del diseño en comparación con otros son menores debido a su sencillez, reduciendo sesgos que estas puedan tener en el modelo.
Tipo de fotocatalizador	TiO_2	Debido a su accesibilidad, bajo costo y por ser uno de los más utilizados y estudiados debido a sus propiedades fisicoquímicas.
Otros parámetros	Tipo de luz: UV	Tipo de fotocatalizador: Nanopartículas TiO2 – P25 en suspensión

Tabla 2. Estructura de la BD.

Columnas	**Rango**	**Unidad**
Nombre del contaminante	-	-
Intensidad de la luz	0.176 - 75	mW/cm^2
Temperatura	20 – 60	°C
Concentración del contaminante	0.13 – 342.47	mg/L
Concentración del catalizador	0.001 – 7.5	g/L
pH inicial	2 – 11	-
pKa ácido	-3 – 15.96	-
pKa básico	-1.99 – 9.27	-
Longitud de onda	172 - 400	nm
Potencia de la lámpara	6 - 1500	W

degradación mediante fotocatálisis heterogénea, resulta complejo modelar y predecir el comportamiento y eficiencia del proceso, dónde además es requerida para su análisis la realización de un gran número de experimentos, donde se evalúa el impacto de las variables que intervienen en él.

Por lo anterior, se presentan en este artículo los resultados de 2 propuesta de modelo de aprendizaje automático, utilizando redes neuronales artificiales (RNA) y bosques aleatorios (BA), para la predicción de la tasa de degradación de contaminantes bajo determinados parámetros experimentales.

2. Metodología

Una vez identificados y delimitados los parámetros que impactan en la eficiencia del proceso, se procedió con la adquisición de los datos que conforman la base de datos, la cual fue utilizada para el entrenamiento y validación de los modelos de aprendizaje automático. La información fue extraída de los resultados experimentales reportados en la literatura, artículos científicos especializados.

En la selección de los artículos se consideraron aquellos que cumplían con los aspectos en la metodología de la experimentación descritos en la Tabla 1, con el fin de

Tabla 3. Análisis exploratorio de los datos.

Parámetro	Correlación	Significancia estadística	R^2	F-Statistic
Temperatura	0.383	1.240×10^{-14}	0.3	3.9×10^{-24}
Intensidad luz UV	0.375	5.650×10^{-06}		
pKa ácido	0.140	4.587×10^{-4}		
Longitud de onda	0.006	3.467×10^{-3}		
pH	0.040	4.098×10^{-03}		
Concentración de catalizador	-0.096	1.106×10^{-02}		
pKa básico	-0.041	1.543×10^{-01}		
Concentración contaminante	-0.095	7.340×10^{-01}		

reducir la variabilidad de los resultados experimentales por este factor. Se detalla en la Tabla 2 la estructura de la base de datos (BD), la cual se conforma de 375 resultados experimentales reportados en 25 artículos de estudio de degradación por fotocatálisis heterogénea. El código necesario fue escrito en lenguaje Python, y ejecutado en el entorno de desarrollo de Google colaboratory, del cual se detallará su implementación en las siguientes secciones.

3. Análisis exploratorio de los datos

En esta etapa se calcularon las siguientes medidas estadísticas, para entender la relación de los datos que conforman la BD, mediante las librerías de `pandas` y `statsmodels`, los resultados obtenidos se resumen en la Tabla 3. Se obtuvo la correlación de las variables independientes con respecto a la tasa de degradación, encontrando que estas presentaban intensidades nulas.

Un algoritmo basado en métodos lineales solo lograría explicar el 30% (R^2) de la variable dependiente, por lo que se decide hacer uso de modelos con la capacidad de adaptarse a la relación no lineal entre las variables, como lo son las redes neuronales y los bosques aleatorios.

Se confirma con la prueba de hipótesis F-statistic que existe alguna relación entre las variables independientes y la dependiente, y con base en los resultados de la significancia estadística, se identificaron las variables con impacto en la variable a predecir (aquellas con un valor < 0.05). Se omiten para el entrenamiento del modelo las variables pKa básico, potencia de la lámpara y temperatura, en el modelo. pKa básico no cumplió con el criterio anterior (significancia estadística).

Se identificó que la correlación entre las variables intensidad de la luz y potencia de la lámpara es alta (correlación = 0.88), se optó por eliminar la variable potencia de lampara, ya que se tienen más registros que reportan la intensidad de la luz, y esta es dependiente de la potencia y distancia de la fuente de luz. Se reporta en la literatura que experimentalmente la variación de la temperatura no es un factor que impacte significativamente la eficiencia del proceso, pero se destaca que el proceso de degradación por fotocatálisis solo ocurre entre los y 20 y 80 °C [4].

Tabla 4. Delimitación de parámetros para el entrenamiento del modelo.

Entrada	• Tipo de contaminante (Molecular Fingerprints) • Concentración de contaminante • pKa ácido • Concentración de fotocatalizador • pH inicial • Intensidad de luz UV • Longitud de onda de la fuente de luz	**Salida**	Tasa de degradación

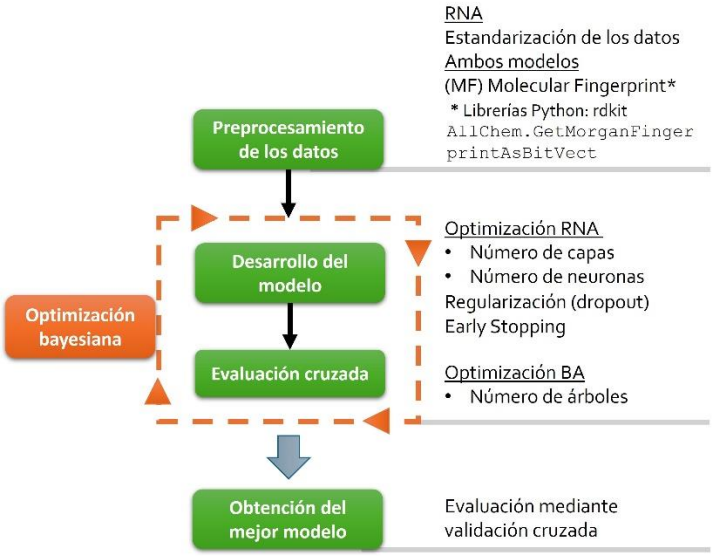

Fig. 1. Metodología para el desarrollo del modelo

4. Desarrollo de modelo

La metodología implementada se resume en la Figura 1, los parámetros considerados para el entrenamiento del modelo se listan en la Tabla 4.

Para que ambos modelos pudieran reconocer al tipo de contaminante, no como un dato categórico, si no con información interpretable de su estructura química, se obtiene del SMILE (por sus siglas en inglés Simplified Molecular Input Line Entry Specification) el vector binario Molecular Fingerprint de cada uno de los compuestos que conforman la BD.

Para el modelo de RNA fue necesario la estandarización de los datos para un adecuado proceso de entrenamiento, en el modelo de BA no es necesaria la

Tabla 5. Resultados de desempeño de ambos modelos.

	RNA	BA
Configuración	• Número de capas = 7 • Número de neuronas = 50 - 250 • Función de activación = ReLu • Optimizador = Adam • Tasa aprendizaje = 0.00094	• Número de árboles = 737
R^2	0.79	0.81
MSE	0.063	0.052
MAE	0.17	0.15
Tiempo de predicción (s)*	0.1733	0.2023
Tiempo de entrenamiento (s) **	9.532	4.978
*Tiempo promedio de 9 predicciones (ejecutado 6 veces) **Tiempo promedio de entrenamiento (Validación cruzada 5 segmentos)		

implementación de esta técnica. Los modelos se implementaron de las librerías `scikit learn` (BA) y `Keras` (RNA).

Los modelos de machine learning generalmente requieren de un proceso de optimización de hiperparámetros para ubicar la configuración con el mejor desempeño de predicción, entre las técnicas de búsqueda más populares se encuentran la búsqueda de cuadrícula, la aleatoria y la bayesiana. Para el modelo de RNA se encontró que no era viable realizar una búsqueda exhaustiva (búsqueda de cuadrícula) para un determinado rango de condiciones, debido a los altos tiempos de procesamiento requeridos, así se decide implementar la optimización bayesiana en ambos modelos, permitiendo encontrar una opción con desempeño destacable.

Se optimiza el número de árboles en BA y el número de capas y neuronas en la RNA (mediante la librería `scikit-optimize BayesSearchCV`), se realizan además pruebas de early stopping y dropout en este modelo. La evaluación se realizó por validación cruzada (5 segmentos), durante el proceso de optimización, y para la evaluación del desempeño de predicción de los modelos, con el fin de reducir al mínimo el sesgo relativo a la división del conjunto de datos.

5. Resultados

Se presenta en la Tabla 5 la comparativa del desempeño de los mejores modelos obtenidos. Considerando los resultados, se observa que BA tiene una pequeña ventaja sobre la predicción en comparación con la RNA, sobre los tiempos, se observa que RNA tarda menos en realizar predicciones con nuevos datos, en comparación de BA, respecto al tiempo de entrenamiento, la RNA tarda significativamente más que BA. En la práctica, el tiempo de entrenamiento también impacta en el tiempo de búsqueda de los mejores hiperparámetros, además, la búsqueda en el modelo de RNA es más amplia

que BA (ya que hay un mayor número de combinaciones posibles), se observó que la optimización de RNA tardo significativamente más que BA.

6. Conclusiones

El análisis y desarrollo propuesto de los modelos que presentaron el mejor desempeño permitió obtener una predicción de la tasa de degradación alrededor del 80% con 7 variables, adaptándose al comportamiento no lineal de las variables con respecto a la variable dependiente.

El trabajo se abordó desde 3 perspectivas: la optimización del desempeño del modelo, el contenido de la base de datos y la selección de los parámetros relevantes que intervienen en el proceso.

En la optimización del desempeño del modelo se busca implementar las herramientas necesarias que permitan que el algoritmo se adapte al patrón de comportamiento de los datos. Se debe buscar que el contenido de la base de datos proporcione la información suficiente y de calidad para distinguir en ella el patrón de comportamiento de estudio.

Por último, se deben considerar todas las variables involucradas en el proceso de estudio necesarias para la caracterización de su comportamiento (todas aquellas que impactan en la variable a predecir). Bajo esta perspectiva se han logrado los resultados obtenidos. La predicción obtenida hasta el momento puede presentar mejoras para disminuir el error de predicción si se añaden datos a la BD, si se considerando la existencia de un posible sesgo por la poca presencia de resultados experimentales expuestos a una variación limitada de las variables de experimentación de cada contaminante. Una BD de calidad es de suma importancia para un adecuado entrenamiento de los modelos de aprendizaje automático, esta no solo se limita a que los datos en ella sean fidedignos, sino que exista en ella información suficiente para identificar patrones adecuadamente. Si bien fue un reto encontrar artículos que reportaran la información requerida, por lo que aumentar la BD requiere la inversión de un tiempo considerable, el trabajo tiene el potencial se asentar las bases y consideraciones de metodología para la adquisición de datos experimentales en laboratorio para la generación e implementación de modelos con aplicaciones similares al presentado.

Referencias

1. Rodríguez, D.J., Serrano, H.A., Delgado, A., Nolasco, D., Saltiel, G.: De residuo a recurso: cambiando paradigmas para intervenciones más inteligentes para la gestión de aguas residuales en América Latina y el Caribe. Banco Mundial, Washington, DC (2020). DOI: 10.1596/33436.
2. Chen, D., Cheng, Y., Zhou, N., Chen, P., Wang, Y., Li, K., Huo, S., Cheng, P., Peng, P., Zhang, R., Wang, L., Liu, H., Liu, Y., Ruan, R.: Photocatalytic degradation of organic pollutants using TiO2-based photocatalysts: a review. Journal of Cleaner Production 268, 121725 (2020). DOI: 10.1016/j.jclepro.2020.121725.
3. Bertagna Silva, D., Buttiglieri, G., Babić, S.: State-of-the-art and current challenges for TiO2/UV-LED photocatalytic degradation of emerging organic micropollutants.

Environmental Science and Pollution Research 28(1), 103–120 (2021). DOI: 10.1007/s11356-020-11125-z.
4. Lee, D.-E., Kim, M.-K., Danish, M., Jo, W.-K.: State-of-the-art review on photocatalysis for efficient wastewater treatment: attractive approach in photocatalyst design and parameters affecting the photocatalytic degradation. Catalysis Communications 183, 106764 (2023). DOI: 10.1016/j.catcom.2023.106764.
5. Malviya, A., Jaspal, D.: Artificial intelligence as an upcoming technology in wastewater treatment: a comprehensive review. Environmental Technology Reviews 10(1), 177–187 (2021). DOI:10.1080/21622515.2021.1913242.