

Índice de la calidad del agua en el Río Santiago (México-Guadalajara)

Anónimo

Abstract El objetivo de este trabajo es crear un Índice de la Calidad del Agua para el Río Santiago en Guadalajara-México mediante modelos de regresión logística y diferentes modelos de machine learning. Se calcularon más de 36 millones de modelos, para encontrar 180 modelos que ayuden a conocer la calidad del agua del río Santiago. Se contruyó un “dashboard” que muestra la información de las muestras recolectadas a lo largo del 2009 al 2022 y la estimación de los parametros de cada uno de los 180 modelos propuestos. Se construyó un modulo de predicción (endpoint) dónde puedan ingresar las muestras recolectadas y conocer si el Río Santiago rebasa las normas de la calidad del agua vigentes.

Palabras clave: Machine learning - Modelos de regresión logísticos - Shiny

##

Found the following errors:

* Missing corresponding author

El objetivo de este trabajo es crear un sencillo Índice de la Calidad del Agua del Río Santiago en Guadalajara-México mediante modelos de regresión logística y diferentes modelos de machine learning para encontrar el óptimo. Un segundo objetivo fué contruir un “[dashboard](#)” que muestre la información de las muestras recolectadas a lo largo del 2009 al 2022 y la estimación del índice en tiempo real con base en las muestras recolectadas. Y un tercer objetivo es construir un modulo de predicción (endpoint) dónde puedan ingresar las muestras recolectadas, en tiempo real y conocer si el Río Santiago rebasa los límites que menciona la norma (Ley Federal de Derechos para la protección de la vida acuática) de la calidad del agua vigente en México.

En Guadalajara (México) cruza la ciudad el Río Santiago, por lo que se ha monitoreado desde el 2008 constantemente hasta el 2022, 38 elementos químicos, que son: Alcalinidad total, Aluminio, Arsénico, Bario, Cadmio, Cloruros totales, Cobre, Conductividad, Cromo, Demanda bioquímica de oxígeno, Demanda química de oxígeno, Dureza total, Fierro, Fluoruros, Fósforo total, Grasas y aceites, Manganeseo, Mercurio, Níquel, Nitrógeno amoniacal, Nitrógeno de nitratos, Nitrógeno de nitritos, Nitrógeno total, Nitrógeno total Kjeldahl, Oxígeno disuelto, Plomo, Potencial de hidrógeno, Sodio, Sólidos disueltos totales, Sólidos sedimentables, Sulfatos, Sulfuros, Sustancias activas al azul de metileno, Temperatura, Temperatura ambiente, Turbiedad y Zinc.

En la actualidad no se cuenta con un dashboard que ayude a la población a conocer cuales son los niveles químicos que presenta el río Santiago en la ciudad de Guadalajara, México. A pesar que se tienen información sobre la calidad del agua del río Santiago no se encuentra de manera de fácil el acceso a dicha información. Tampoco se cuenta con un índice de la calidad del agua en tiempo real sobre el río Santiago.

La norma (Ley Federal de Derechos para la protección de la vida acuática) señala ciertos límites máximos permisibles de algunos elementos químicos que debe tener en el Río Santiago. El objetivo es crear un índice la calidad del agua con base en *nueve variables dependientes*, con un nivel de medición dicotómicas. Con base en la información de los límites que debe tener el río Santiago, se crearon dichas variables dependientes de la siguiente manera: si superaban los límites permisibles se consideraba como uno y cero en otro caso. Los límites de los valores permisibles por la norma que se tomaron fueron los siguientes: Fluoruros ≤ 1.5 , Nitrógeno de nitritos ≤ 1 , Aluminio ≤ 0.2000 , Cadmio ≤ 0.005 , Cromo ≤ 0.05 , Fierro ≤ 0.3 , Manganeseo ≤ 0.15 , Mercurio ≤ 0.001 , y Plomo ≤ 0.01 .

En la primera etapa, con estas 38 mediciones de los elementos químicos que se obtuvieron en cada lugar muestreado del río Santiago, se consideró cómo variables independientes dichos elementos químicos.

El objetivo en esta etapa es encontrar una combinación de alguno de estos 38 elementos químicos (variables independientes) que aporte la mayor información que permita predecir si se supera los límites máximos permisibles de los elementos químicos. Comenzamos aplicando modelos de regresión logísticas, calculando la pseudo r cuadrada (R^2), el criterio de información de Akaike (AIC), el criterio de información bayesiano (BIC) y la deviance.

Se calculó un total de **36,537,426** de modelos de regresión logística. Para cada una de las nueve variables independientes se calcularon 5,706 modelos de regresión logística con dos variables independientes; 64,890 modelos de regresión logística con tres variables independientes; 537,285 modelos de regresión logística con cuatro variables independientes y 3,451,833 modelos de regresión logística con cinco variables independientes. Se decidió no seguir calculando más modelos ya que la pseudo r cuadrada alcanaban valores cercanos a uno en todos estos modelos calculados.

En la segunda etapa, después de calcular todos los modelos de regresión logística se decidió trabajar con los mejores 180 modelos. Aplicamos métodos de machine learning para determinar que función es mejor. El objetivo en esta segunda

etapa es evaluar estos 180 modelos de manera eficiente usando los métodos de carrera. Dividimos los datos usando la proporción del 80 por ciento de los datos para entrenamiento y el 20 por ciento para pruebas y volvimos a muestrear el conjunto de entrenamiento usando cinco repeticiones de validación cruzada. Algunos modelos de machine learning requieren preprocesamiento de los datos: que estén en las mismas unidades.

Para evaluar su rendimiento de cada modelo se utilizó las funciones estándar de ajuste o remuestreo. Se garantizó que cada ejecución ocupará los mismos números aleatorios.

Para evaluar los rendimientos de los resultados de cada modelo se ocupó el **RMSE**, finalmente el objetivo es encontrar la mejor combinación de parámetros de ajuste de cada modelo.

En conclusión, el aporte a este trabajo son: se realizó una propuesta la construir un índice de la calidad del agua del río Santiago de manera sencilla. Segundo, se puede calcular en tiempo real los valores del índice de la calidad del agua del río Santiago. Y tercero, se contruyó un dashboardo dónde se puede observar la información recolectada sobre el monitoreo del río Santiago.