PRONÓSTICO DE LA CALIDAD ESPACIO TEMPORAL DEL AGUA EN EL ESTADO DE GEORGIA ESTADOS UNIDOS

N. Urueta Peñata[1]

1. Aspirante a Maestrante en Estadística Aplicada. Universidad Tecnológica de Bolívar, Instituto Nacional de Medicamentos y Alimentos Invima

El presente estudio tiene por objetivo lograr encontrar un modelo que permita pronosticar la calidad del agua de acuerdo a varias variables físico-químicas en el estado de Georgia Estadios unidos. La base de datos a utilizar es proporcionada por el servicio Geológico de los Estados Unidos. Contiene información de la calidad del agua medida en 37 lugares distinto del estado de Georgia y en base al Potencial de Hidrogeno, este atributo será la variable para pronosticar en función de 9 atributos tomados de una muestra de 423 días. Para los 37 lugares se tomaros medidas diarias del volumen de oxígeno disuelto, la temperatura y la conductancia específica, del total de las mediciones por variable se sacaron diariamente los valores mínimos, promedios y máximos, generando así los 9 atributos.Para la variable de respuesta (PH) de las 37 medidas se tomó como medida diariamente la mediana. Los datos estaban almacenados en un archivo .MAT,(de Matlab) que al ser extraído en python generaban un archivo de tipo DICT (Diccionario). De este se extrae para consolidar la base de datos la información de las llaves o carpetas asociadas a los datos de los atributos predictivos, los datos de la variable de respuesta y los nombres de las variables. Las metodologías utilizadas para el tratamiento de la información obtenida y la emisión de resultados fueron las herramientas de Machine Learning con regresión lineal múltiple y Machine Learning con Random Forest. Como resultado se obtuvo un valor de RMSE de 0,0045 con la metodología de Machine Learning con regresión lineal múltiple y 0,0017 para el método con Random Forest, logrando este último establecer una mejor regresión con un ajuste del 99,08% de los datos contra un 93,60% del modelo con regresión lineal. Para mejor apreciación del ajuste ver la figura presente en este resumen. Al lograr determinar que este tipo de modelos es muy buenos para estas variables físico-químicas para trabajo futuro ya que el país de Estados Unidos presenta varios sus horarios se recomendaría aplicar esta metodología con estados que abarquen todos los uso horarios siendo estos una variable y el miso estado otra variable.

Figura 1. Ajuste de datos mediante los Métodos de regresión lineal con Machine Learning