Instituto Tecnológico y de Estudios Superiores de Monterrey

Campus Monterrey

Aplicación de métodos multivariados en ciencia de datos (Gpo 101)

Impacto de factores no antropogénicos en los niveles de contaminación del aire

Equipo 4 | Integrantes:

Daniel Eduardo Arana Bodart | A01741202

Isis Yaneth Malfavón Díaz | A01705838

Santiago Juarez Roaro | A01705439

Ericka Sofía Rodríguez Sanchez | A01571463

Reporte individual de:

Alfredo André Durán Treviño | A01286222

Profesores:

Dra. Blanca Rosa Ruiz Hernandez

&

Mtro. Rodolfo Fernández de Lara Hadad

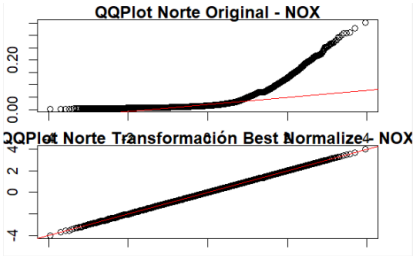
7 de septiembre del 2024

Nuestro proyecto fue enfocado en la pregunta: ¿Existe un impacto significativo de los factores no antropogénicos en los niveles de contaminantes del aire? Esta propuesta fue hecha en equipo después de varios intentos de hacer un objetivo, ya que en cada entrega cambiaba de propósito, por ejemplo, la segunda entrega estaba enfocada a los niveles de calidad del aire, por lo que habíamos hecho una nueva variable que nos dijera si la calidad del aire era buena o no dependiendo de la información en dicha hora, viendo así si con las características meteorológicas se podría predecir la calidad del aire, lamentablemente este propósito no funcionó, por lo que después buscamos en los contaminantes los que se vean más afectado por las características meteorológicas, encontrando que el ozono (O3) es la que tiene una mayor relación, aunque seguía sin ser significativa, pero decidimos dar ese enfoque en el ozono.

Primero analizamos los datos que teníamos, haciendo una limpieza de los datos, ya que de la base de datos que usamos para nuestros modelos finales fue la base de datos Norte 2023-2024, de las cuales un 7.94% de los datos eran nulos, por lo que era necesario hacer una limpieza de estos, para hacerlo usamos el método KNN (K-nearest neighbors), esto tras ser una serie de tiempo, este método usa los valores cercanos para sacar un estimado de los valores vacíos. Tras la limpieza de los datos nulos, se escalaron usando Standard Scaler de Scikit Learn, para de esta manera se puedan usar estos datos en el modelo. Tras esto surgió un problema, la normalidad de los datos, ya que tras ver las gráficas, se mostraba que no había normalidad en ninguna variable, y esta característica era necesaria para nuestros modelos, por lo que tuvimos que recurrir a una transformación para que haya normalidad, lo que fue difícil, ya que no funcionaban las transformaciones que usábamos, hasta que usamos la transformación de quantiles ordenados de bestNormalize, que usaba la siguiente fórmula:

$$g(x^* \mid \mathbf{x}) = \begin{cases} f(x^*) & \text{if } x^* \in \{\mathbf{x}\} \\ \dfrac{f(x_u) - f(x_l)}{x_u - x_l} & \text{if } x^* \notin \{\mathbf{x}\} \text{ and } \min \mathbf{x} < x^* < \max \mathbf{x} \\ r(x^*; \mathbf{x}) & \text{if } x^* < \min \mathbf{x} \text{ or } x^* > \max \mathbf{x} \end{cases}$$

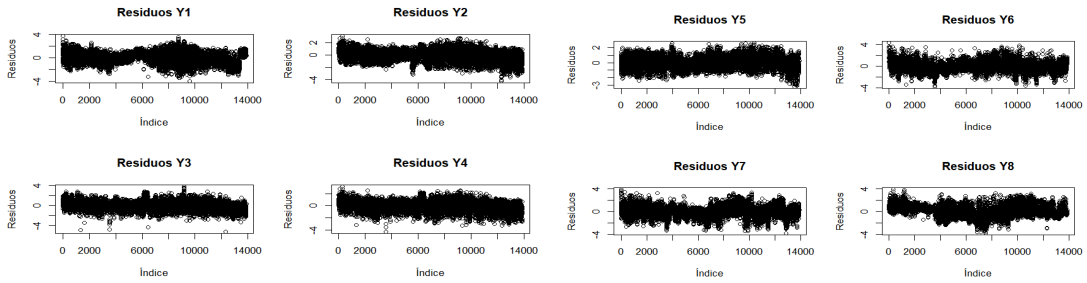
Usando condiciones para una fórmula diferente para cada dato, ya con esto se logró que nuestras



variables tuvieran una distribución normal:

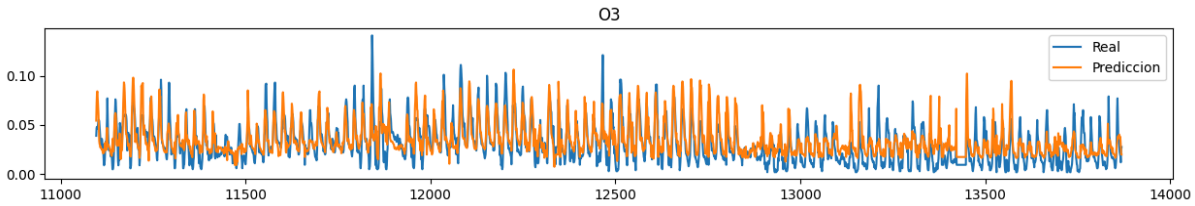
Como se ve en el qqplot anterior. Las variables pudieron ser normalizadas para usar un modelo de regresión lineal multivariada, para esto no se usaron las variables SR RAINF ni WDV, tras sacar estos modelos, vimos las R cuadrada de cada uno y al parecer el ozono era el que mejor se ajustaba, con un valor de 0.464, que nos es un valor alto, pero nos fue interesante que fuera el mejor de los modelos que sacamos.

Durante la validación del modelo usamos diagramas de dispersión de los residuos para demostrar el supuesto de homocedasticidad, viendo los gráficos de dispersión de los residuos, considerando que si parece haber un patrón en los residuos no se cumple el supuesto de homocedasticidad, concluyendo en que se cumple el supuesto.



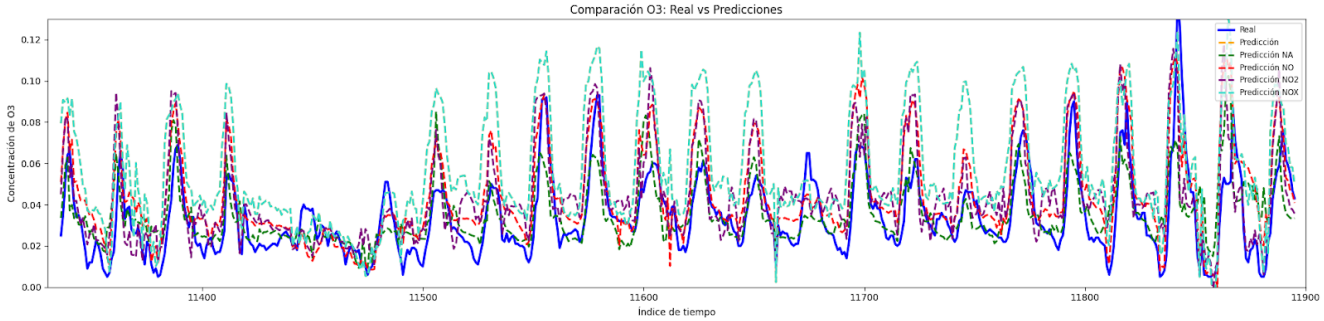
También revisamos la normalidad en los residuos, concluyendo que si hay normalidad. Finalmente vimos la linealidad de las variables, usamos correlación canónica con un valor de 0.759, viendo que tiene una linealidad aceptable. Pero desgraciadamente no se cumplió el supuesto de multicolinealidad, la determinante de la matriz de correlación era de 0.0027, menor a 0.01, por lo que no era recomendada.

Usamos un segundo modelo para el Ozono, de redes neuronales, para así poder tomar en cuenta la dependencia de los datos, usando una memoria de corto plazo LSTM para que funcione con la serie de tiempo, se usó una capa densa de 64 neuronas con activación relu, una capa dropout, otra capa densa de 32 neuronas con relu, y una capa de salida de una neurona, finalmente una función de pérdida del error medio absoluto.



El resultado de el modelo fue bastante aceptable, con un error medio absoluto de 0.01014, viendo la gráfica, vemos que le es complicado para el modelo predecir los picos más grandes, esto se

hizo con factores no antropogénicos, luego intentamos hacer el modelo con otros factores, como los contaminantes NO, NO2 y NOX, y los resultados variaron de manera similar.
Mostramos el gráfico con los modelos que se hicieron:



La conclusión que obtuvimos para la pregunta es que si se puede predecir significativamente el nivel de un contaminante solamente con factores no antropogénicos, ya que aprendimos que si hay una relación importante entre estos factores meteorológicos (no antropogénicos) con este contaminante.