PREDICCIÓN DE LA CONTAMINACIÓN DEL AIRE POR PARTÍCULAS (PM10) MEDIANTE MACHINE LEARNING K-NEAREST NEIGHBORS

Forecasting particulate air pollution (PM10) using a Machine Learning K-Nearest Neighbors

Valentina Botero Mármol
Universidad Libre, Bogotá, Colombia,
valentina-boterom@unilibre.edu.co

Ingrid Dayana Sánchez Martin
Universidad Libre, Bogotá, Colombia,
ingridd-sanchezm@unilibre.edu.co

RESUMEN

La contaminación del aire es un tema que ha venido tomando cada vez más relevancia, puesto que ha aumentado debido a la urbanización, industrialización, entre otras; además, es un grave problema que afecta la salud de las personas y el medio ambiente. Es causada por la emisión de gases tóxicos derivados ya sea por actividades industriales y/o antropogénicas. La calidad del aire se ha convertido en una preocupación global, y se requieren medidas para reducir las emisiones y mejorar la calidad del aire en todo el mundo. Esta investigación permitió predecir los valores de material particulado PM10 en la localidad de Kennedy en Bogotá con ayuda del algoritmo de Machine Learning K-Nearest Neighbors. Además, se obtuvieron resultados alentadores frente a la generalización de los datos de prueba, lo que demostró una buena confiabilidad ante sus predicciones.

Palabras clave: Contaminación del aire; predicción; material particulado (PM10); Kennedy; KNN.

ABSTRACT

Currently, air pollution is an issue that has become increasingly relevant, since it has increased due to urbanization, industrialization, among others; Furthermore, it is a serious problem that affects people's health and the environment. It is caused by the emission of toxic gases caused either by industrial and/or anthropogenic activities. Air quality has become a global concern, and measures are required to reduce emissions and improve air quality around the world. This research made it possible to predict the values of PM10 particulate matter in the town of Kennedy in Bogotá with the help of the K-Nearest Neighbors Machine Learning algorithm, in addition, obtaining encouraging results regarding the generalization of the test data, demonstrating good reliability regarding its predictions.

Keywords: Air pollution; Prediction; Particulate Matter (PM10); Kennedy; KNN.

1. INTRODUCCIÓN

La contaminación del aire se ha convertido en un tema crucial en la sociedad moderna, por cuanto ejerce una influencia significativa tanto en la salud humana como en las actividades diarias. Para contrarrestar las graves consecuencias de los episodios de contaminación, las autoridades han establecido redes de monitoreo de la calidad del aire que proporcionan información en tiempo real sobre el índice de calidad del aire. A través de diversas técnicas, como métodos estadísticos, inteligencia artificial y técnicas de extracción de datos, los conjuntos de datos registrados y guardados en almacenes de datos pueden analizarse para prever futuros valores de los parámetros más críticos de contaminación del aire, así como del índice de calidad del aire.

Uno de los métodos más destacados es el K vecinos más cercanos (KNN), que se basa en calcular la distancia entre una nueva instancia y sus K vecinos más cercanos para clasificarla. Este enfoque también se puede emplear para predecir el valor de un parámetro específico. En este artículo se presenta un experimento detallado, realizado con el objetivo de explorar las aplicaciones específicas de esta técnica en el análisis de la calidad del aire, centrado en la estación de monitoreo de Kennedy y específicamente en la variable PM10.

2. REVISIÓN DE LITERATURA

Se ha realizado una variedad de investigaciones que han centrado su atención en analizar a partir de algoritmos como "KNearest Neighbor" los diferentes contaminantes del aire, entre los cuales encontramos el PM10. Está el caso de la investigación "Spatial assessment of PM10 hotspots using Random Forest, K-Nearest Neighbour and Naïve Bayes" donde se utilizaron diferentes algoritmos entre los cuales se hallaba el K-Nearest Neighbors, implementados con el fin de predecir puntos críticos de peligro espacial de PM10, esto con el fin de garantizar una preparación adecuada para la gestión de la calidad del aire y así minimizar los efectos. Los datos modelados de PM10 fueron adquiridos entre 2012 - 2016, y los resultados obtenidos revelaron una buena predicción, la cual presentó una precisión general para el K-Nearest Neighbour (KNN) del 96% [1].

Otra investigación es "Air Quality Index Prediction using K-Nearest Neighbor Technique", en la cual se ha utilizado este algoritmo para la predicción de la calidad del aire; el KNN predice series temporales caóticas de concentraciones de contaminantes atmosféricos, para pronosticar con 8 horas de anticipación las emisiones de SO2, CO, NO2, NO y O3 y niveles máximos de CO. Por esta razón, en este artículo fue implementado para predecir la calidad del aire, utilizando una base de datos con valores registrados para el dióxido de azufre, óxidos de nitrógeno, monóxido de carbono y ozono con ayuda de WEKA, un software utilizado para minería de datos. Los resultados de esta investigación fueron muy positivos, ya que en 19 de los 29 casos el error de predicción fue del 0% [2]

3. ÁREA DE ESTUDIO Y METODOLOGÍA

La contaminación del aire por partículas suspendidas (PM10) es una problemática grave en muchas áreas urbanas, incluido en la localidad de Kennedy en Bogotá, Colombia. Las partículas PM10 son partículas inhalables con un diámetro de 10 micrómetros o menos, lo que las hace lo suficientemente pequeñas como para ser inhaladas y retenidas en los pulmones, lo cual causa problemas respiratorios.

ISSN 2322-8415 / Año 2024 / No. 23 - Vol. 13 / Páginas 57-64 Universidad Libre - Facultad de Ingeniería

En áreas urbanas como Kennedy, la contaminación del aire por PM10 puede provenir de diversas fuentes, como el tráfico vehicular, las actividades industriales, la construcción, la quema de biomasa y el polvo proveniente de suelos desprotegidos. Estas partículas pueden contener sustancias nocivas, como metales pesados, compuestos orgánicos y sustancias químicas tóxicas, que representan un riesgo para la salud humana, especialmente para niños, personas mayores y personas con afecciones respiratorias preexistentes.

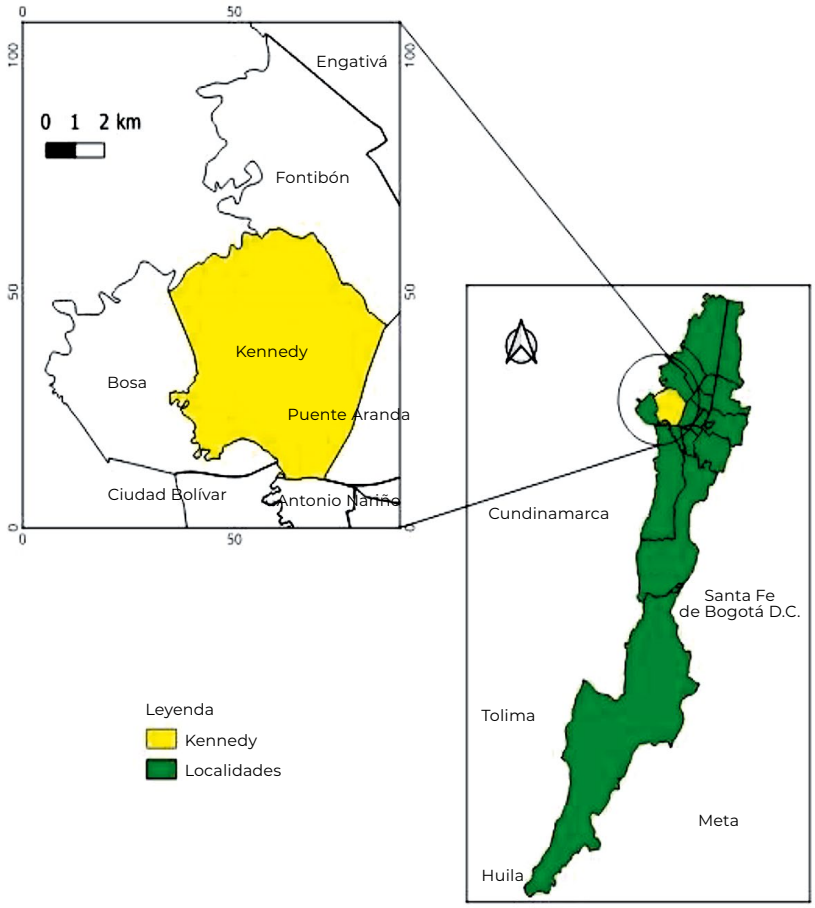
3.1.1 Descripción del conjunto de datos

Los datos cronológicos implementados para esta investigación fueron obtenidos de una estación de calidad de aire de la localidad de Kennedy, para estudiar el material particulado PM10, haciendo uso del algoritmo K-Nearest Neighbors, que permite hacer predicciones por medio técnicas de estimación adecuadas. La práctica para este algoritmo suele ser dividir los datos observados durante un período de tiempo en dos grupos, el primero de los cuales se utiliza para obtener las estimaciones del segundo grupo.

4. MODELOS

Las PM10 se pueden definir como aquellas partículas sólidas o líquidas de polvo, cenizas, hollín, partículas metálicas, cemento o polen, dispersas en la atmósfera, y cuyo diámetro varía entre 2,5 y 10 µm (1 micrómetro corresponde la milésima parte de 1 milímetro) [3].

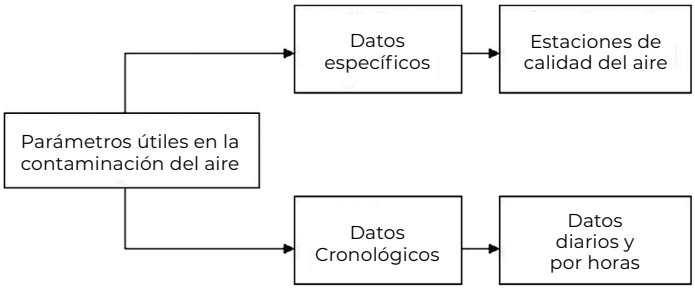
Fig. 1: Ubicación de la zona de estudio



Este tipo de partículas PM10 pueden generarse en la atmósfera tanto por causas naturales como por efecto del hombre, es decir por causas antropogénicas [4]; además, una exposición prolongada o repetitiva a las PM10 puede provocar efectos nocivos en el sistema respiratorio de las personas.

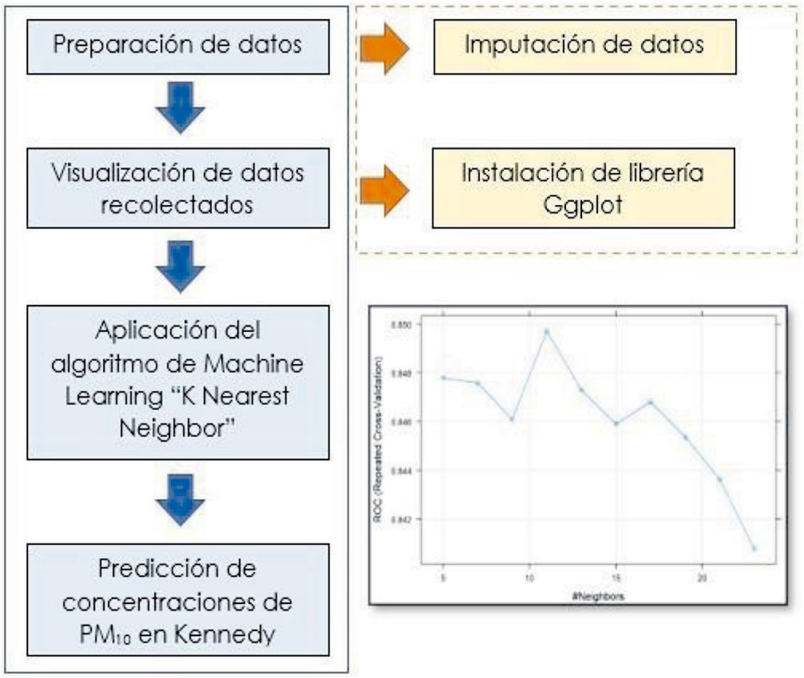
Este informe describe las emisiones de material particulado de 10 micras. Esta investigación se realizó en la localidad de Kennedy porque los estudios distritales señalan que, en esta localidad, que es la segunda más grande de Bogotá, confluyen muchos problemas de carácter ambiental, dentro de los cuales está la contaminación atmosférica. Los informes anuales de la calidad de aire de Bogotá señalan que esta localidad es de las más contaminadas de la ciudad [5].

Fig. 2: Datos de contaminación en la investigación



Teniendo en cuenta los datos recolectados a partir de las estaciones de calidad de aire en la localidad de Kennedy, se propone una ruta de trabajo para la investigación, tal como se muestra en la Figura 3.

Fig. 3: Modelo de predicción propuesto para la predicción de material particulado PM10



ISSN 2322-8415 / Año 2024 / No. 23 - Vol. 13 / Páginas 57-64 Universidad Libre - Facultad de Ingeniería

Debido a que los datos recolectados inicialmente presentan datos faltantes, cuyos patrones representan relaciones matemáticas genéricas entre los datos observados y los ausentes, es necesario realizar una imputación de datos que consiste en el remplazo de los valores faltantes por los valores de la media, mediana o moda, y en general es el de uso más común por su facilidad de implementación; en este caso, se aplica la imputación por el valor de la media, para continuar con la aplicación del K-Nearest neighbors, un algoritmo supervisado de Machine Learning, el cual se usa para clasificar nuevas muestras (valores discretos) o para predecir (regresión, valores continuos). Al ser un método sencillo, es ideal para introducirse en el mundo del aprendizaje automático. Sirve esencialmente para clasificar valores buscando los puntos de datos "más similares" (por cercanía) aprendidos en la etapa de entrenamiento [6].

A diferencia de K-means, que es un algoritmo de agrupamiento no supervisado donde "K" se refiere al número de grupos que se desean crear, en K-Nearest Neighbors, "K" representa la cantidad de puntos cercanos que se consideran al clasificar los datos en "n" grupos previamente conocidos. K-NN es un algoritmo supervisado, lo que significa que se basa en etiquetas de datos existentes para tomar decisiones de clasificación.

4.1 K-Nearest neighbors

K-nearest neighbors es un algoritmo de aprendizaje automático supervisado no paramétrico que se utiliza para clasificación y regresión. Calcula la similitud entre observaciones para lo cual se basa en una función de distancia (normalmente euclidiana) y se prefiere cuando los datos son continuos. Un nuevo punto X se clasifica en función de sus K vecinos más cercanos por distancia. El supuesto aquí es que las nuevas observaciones se comportarán como sus vecinos más cercanos [7].

5. CRITERIO DE DESEMPEÑO

Para evaluar el rendimiento del modelo Knearest neighbors, se requiere el uso de evaluaciones estadísticas, algunas de las cuales son: error absoluto medio (MAE), error cuadrático medio de raíz (RMSE), error cuadrático medio (MSE) y el coeficiente de determinación (R2). Las fórmulas se describen a continuación:

Error absoluto medio (MAE):

$$MAE = \frac{\sum_{i=1}^{m} |x_i - \hat{x}_i|}{m} \qquad (1)$$

Donde m es el número de observaciones, $\hat{x}_i$ es el valor predicho, y x_i es el valor real.

Error absoluto medio (MAE):

$$MSE = \frac{1}{m}\sum_{j=1}^{m} (x_i - \hat{x}_i)^2 \qquad (2)$$

Donde $\hat{x}_i$ es el pronóstico, m es el número de veces diferentes para las cuales se pronostica la variable y x_i es el valor real de la cantidad que se pronostica.

Coeficiente de determinación (R2):

$$R^2 = \left[\frac{1}{M}\frac{\sum_{j=1}^{m}(A_j - \bar{A})(B_j - \bar{B})}{\sigma_a \sigma_b}\right]^2 \qquad (3)$$

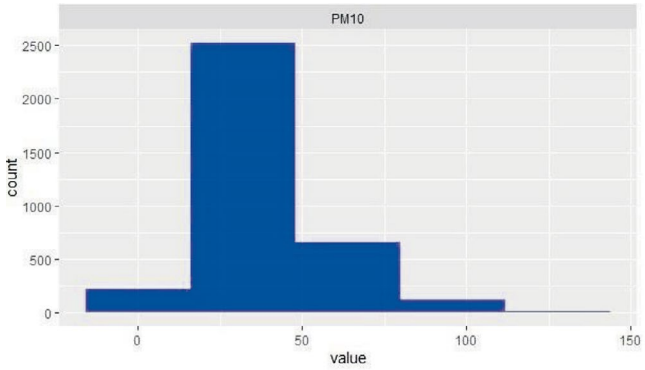
Donde M es el número de observaciones, σ_b es la desviación estándar de la observación, σ_a es la desviación estándar del cál-

culo, Bj es el parámetro observado, B es la media del parámetro observado, Aj es el parámetro calculado y A es la media del parámetro calculado [8]

6. RESULTADOS Y DISCUSIONES

A partir de los datos recolectados e imputados, se pudo obtener una primera visualización de la información, y se representaron mediante un histograma, tal como se muestra en la Figura 4.

Fig. 4: Histograma con datos recolectados



Cabe destacar que para la implementación del modelo KNN fue necesario estandarizar los datos obtenidos previamente, con el fin de representarlos en una escala de 0 a 1. Seguido a eso, se procedió a la división de la información para clasificarlos en datos de entrenamiento y prueba. A continuación, se muestran los diagramas para cada uno de los datos.

Fig. 5 Datos de entrenamiento

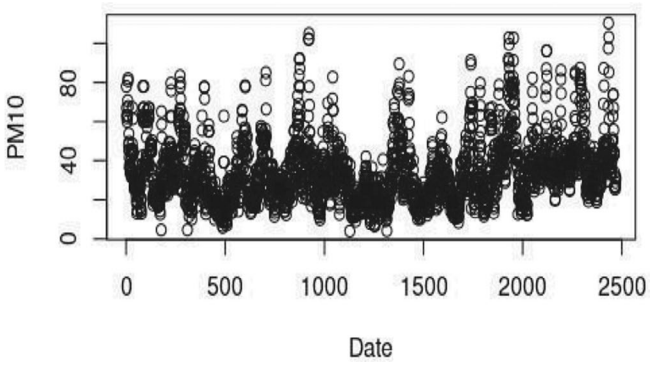
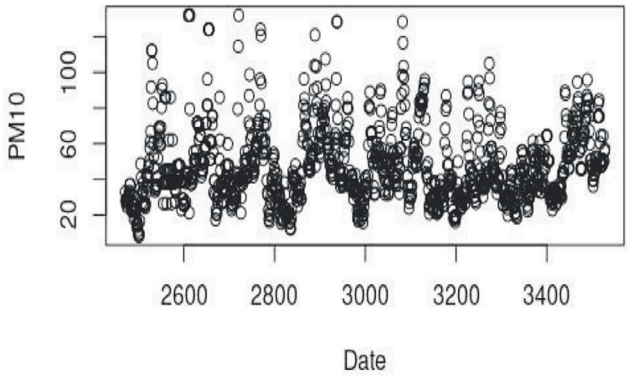
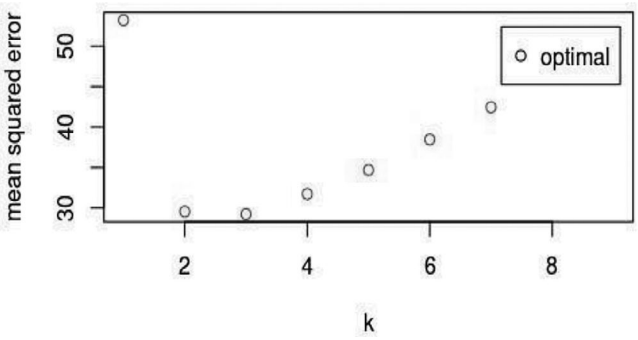


Fig. 6 Datos de prueba



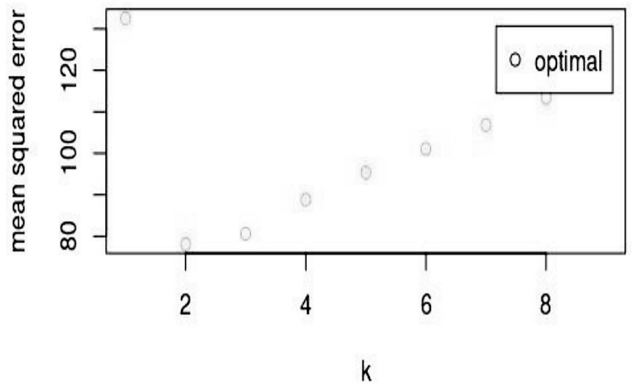
Luego de haber aplicado el modelo KNN a partir de los datos de entrenamiento, se obtuvo como resultado:

Fig. 7 Modelo KNN para datos de entrenamiento



De la misma forma, se realizó el procedimiento para los datos de prueba, y se obtuvo como resultado el siguiente diagrama:

Fig. 8 Modelo KNN datos de prueba



ISSN 2322-8415 / Año 2024 / No. 23 - Vol. 13 / Páginas 57-64 Universidad Libre - Facultad de Ingeniería

Finalmente, se aplica la predicción tanto para los datos de entrenamiento como los de prueba, tal como se muestra en las Figuras 9 y 10, respectivamente.

Fig. 9 Predicción para datos de entrenamiento

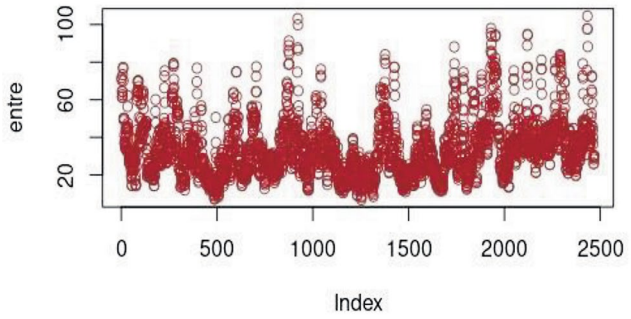
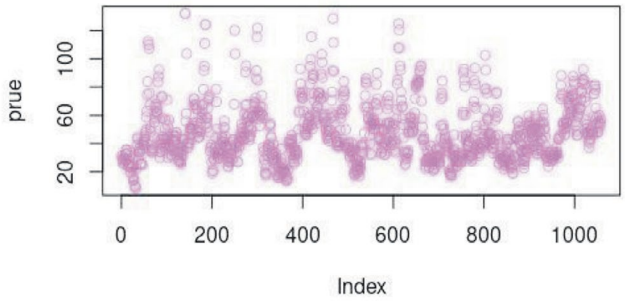


Fig. 10 Predicción datos de prueba



A partir de lo anterior, se logran establecer los criterios de desempeño para los datos, los cuales se reflejan mediante la Tabla 1.

Tabla 1. Criterios de desempeño

Datos de entrenamiento			**Datos de prueba**		
R2	MAE	MSE	R2	MAE	MSE
1	3,747	29,224	1	5,141	78,073

7. CONCLUSIONES

- En conclusión, la similitud en los gráficos de rendimiento entre los conjuntos de entrenamiento y prueba sugiere que el modelo KNN está logrando un buen equilibrio entre la capacidad de ajuste y la capacidad de generalización, ya que la mayor concentración de PM10 tanto para los datos de prueba como para los de entrenamiento se encuentran entre 20 y 60.

- En definitiva, este modelo de KNN ha demostrado una buena capacidad de generalización en los datos de prueba, lo que demuestra que las predicciones para PM10 pueden ser más confiables y aplicables para las predicciones del comportamiento de este contaminante.

- Así pues, es evidente la existencia de datos atípicos, ya que el error cuadrático medio (MSE) en el caso de datos de prueba presentó un alto valor de 78,073, lo cual demostró la existencia de una variedad considerable de errores individuales.

- Para finalizar, los valores de error absoluto medio (MAE) tanto para los datos de entrenamiento como para los de prueba fueron respectivamente de 3,747 y 5,141, lo que indica un buen modelo de predicciones, ya que se contaba con una pequeña magnitud.

REFERENCIAS BIBLIOGRÁFICAS

[1] *A. Tella, A. L. Balogun, N. Adebisi, and S. Abdullah,* "Spatial assessment of PM10 hotspots using Random Forest, K-Nearest Neighbour and Naïve Bayes," Atmos Pollut Res, vol. 12, no. 10, p. 101202, Oct. 2021, doi: 10.1016/J.APR.2021.101202.

[2] *E. Dragomir,* "Air Quality Index Prediction using K-Nearest Neighbor Technique," 2010.

[3] *"Partículas PM10 | PRTR España."* Accessed: nov. 04, 2023. [Online]. Available: https://prtres.es/Partículas-PM10,15673,11,2007.html

[4] *"Las partículas PM10 y cómo se generan."* Accessed: Nov. 04, 2023. [Online]. Available: https://www.eurofins- environment.es/es/particulas-pm10-y-como-se- generan/

[5] *C. B. Gil et al.,* "Contaminación del aire y enfermedades respiratorias, un estudio en la localidad de Kennedy," Sep. 2020, doi: 10.57998/BDIGITAL.HANDLE.001.3530.

[6] *"Algoritmo k-Nearest Neighbor | Aprende Machine Learning."* Accessed: Nov. 04, 2023. [Online]. Available: https://www.aprendemachinelearning.com/clasificar-con-k-nearest-neighbor- ejemplo-en-python/

[7] *"RPubs – K Nearest Neighbors."* Accessed: Nov. 04, 2023. [Online]. Available: https://rpubs.com/awanindra01/knn

[8] *"(PDF) Forecasting Air Pollution Particulate Matter (PM 2.5)* Using Machine Learning Regression Models." Accessed: Nov. 04, 2023. [Online]. Available: https://www.researchgate.net/publication/338 254310_Forecasting_Air_Pollution_Particulate_Matter_PM_25_Using_Machine_Learning_Regression_Models