PONTIFICIA UNIVERSIDAD CATÓLICA DEL PERÚ

ESCUELA DE POSGRADO

Implementación de un modelo algorítmico para la estimación del nivel de concentración de contaminante PM2,5 en zonas urbana

TRABAJO DE INVESTIGACIÓN PARA OPTAR EL GRADO ACADÉMICO DE MAGÍSTER EN INFORMÁTICA CON MENCIÓN EN CIENCIAS DE LA COMPUTACIÓN

Autor: Irvin Rosendo Vargas Campos

Asesor: Dr. Edwin Rafael Villanueva Talavera

Julio, 2020

Resumen

Según la Organización Mundial de la Salud (OMS), la mala calidad del aire provoca 1 de cada 10 muertes globalmente, 7 millones de personas fallecen al año debido a enfermedades causadas por la contaminación, además la mala calidad del aire es un factor contribuyente al cambio climático, específicamente al calentamiento global. En Perú, se debe cumplir los Estándares de Calidad Ambiental (ECAs) establecidos por el Ministerio del Ambiente y supervisados por el Organismo de Evaluación y Fiscalización Ambiental (OEFA); no obstante, cumplir esta tarea se ve dificultada por la baja cantidad de estaciones de medición. Debido a ello, el presente proyecto propone estudiar diferentes estrategias de ingeniería de características y modelos de aprendizaje de máquina que puedan estimar el nivel de contaminación de aire en zonas urbanas no censadas. Para ello, se usó datos de contaminantes y variables meteorológicas recolectados por una red de monitoreo en la ciudad de Beijing, China. Se obtuvo como resultado que el modelo *Linear Regression* entrenado con los datasets de contaminante $PM_{2,5}$ de las 5 estaciones más cercanas al punto de predicción y normalizados mediante una adaptación de *Inverse Distance Weighting* presentó mejor capacidad de estimación. Por otro lado, los modelos *LightGBM* y *XGBoost* presentaron resultados un poco inferiores, pero eran más robustos, pues su capacidad de estimación se mantenía estable a pesar de la modificación de la cantidad de estaciones usadas para el entrenamiento de los modelos. Como trabajo futuro, se pretende usar y adaptar los modelos estudiados en esta investigación en las zonas urbanas de Lima, Perú.

Índice

1. Introducción **3**

2. Trabajos relacionados **4**

3. Metodología **8**

4. Experimentación **9**

4.1. Línea base 9

4.2. Modelos de aprendizaje de máquina 10

4.2.1. Procesamiento de datos 11

4.2.2. Entrenamiento y prueba de modelos 12

5. **Resultados y Discusiones** **13**

6. **Conclusiones** **17**

7. **Trabajos futuros** **18**

8. Referencias **19**

1. Introducción

El aire es el recurso más valioso del planeta y hoy en día es amenazado por los altos niveles de contaminación. Según la Organización Mundial de la Salud (OMS), la mala calidad del aire provoca 1 de cada 10 muertes globalmente, 7 millones de personas fallecen al año debido a enfermedades causadas por la contaminación [16]; además, también es un factor contribuyente al cambio climático, específicamente al calentamiento global debido al aumento de concentraciones de los gases de efecto invernadero. En Perú, el aire es contaminado a altos niveles, comparado con otros países de Latinoamérica [7]. Entre los efectos a corto plazo de estar expuestos a ambientes con elevada contaminación están: tos, dolor de pecho, dolor de cabeza, náuseas, bronquitis y neumonía. Los efectos a largo plazo incluyen cáncer de pulmón, enfermedades cardiovasculares y respiratorias y alergias [9].

El monitoreo de calidad del aire en Perú solo se realiza en Lima, con una red de 10 estaciones de SENAMHI que monitorean una área de 3000 km^2 para 10 millones de habitantes[3]. En el caso de Perú, se debe cumplir los Estándares de Calidad Ambiental (ECAs) establecidos por el Ministerio del Ambiente y supervisados por el Organismo de Evaluación y Fiscalización Ambiental (OEFA); no obstante, cumplir esta tarea se ve dificultada por la baja cantidad de estaciones de medición. Esta escasez se debe a los altos costos de dichas estaciones, así como sus altos costos de mantenimiento y operación.

Ha habido pocos intentos de desarrollar modelos computacionales capaces de producir mapas de contaminación ambiental en Perú con una resolución espacial y temporal que sea de utilidad para las autoridades y los ciudadanos. Entre esos intentos, se encuentran los recientes trabajos de Sánchez-Ccoyllo et al. (2018) [8] y Reátegui et al. (2018) [7], que implementaron modelos numéricos de contaminación WRF-CHEM para predecir material particulado $PM_{2,5}$ en Lima. La resolución espacial fue de 5x5 km y la resolución temporal fue de hora en hora. Los resultados, validados con mediciones de las estaciones del SENAMHI en dos meses del 2016, mostraron la dificultad de tener estimaciones cercanas a los valores reales. Dichas dificultades fueron atribuidas a las incertezas en el inventario de emisiones y las simplificaciones paramétricas que se tuvieron que realizar. Estas dificultades son usuales en los modelos numéricos, los cuales necesitan especificaciones precisas de los distintos parámetros y las condiciones iniciales y de frontera para generar simulaciones realistas. Los modelos numéricos son también altamente demandantes de recursos computacionales, ya que necesitan resolver complejos sistemas de ecuaciones, por lo que su escalabilidad a resoluciones espaciales y temporales mas altas es comúnmente problemática. Ello va en contraposición con algunos estudios en Lima que indican que los gradientes de contaminación relevantes

[3]ver http://www.senamhi.gob.pe/?p=monitoreodecalidad-de-aire

para la salud de las personas se dan justamente a escalas sub-kilométricas [1].

Ante esa necesidad, el Grupo de Inteligencia Artificial PUCP (IA-PUCP) junto con la Universidad Católica San Pablo y apoyados por el Fondecyt-BM vienen implementando un sistema de monitoreo de calidad del aire con sensores de bajo costo para zonas urbanas. Sin embargo, además del desarrollo de la red de medición, uno de los componentes claves es desarrollar modelos algorítmicos capaces de inferir niveles de contaminación en puntos geográficos no censados por los módulos de medición a fin de tener un sistema de monitoreo funcional que produzca mapas de contaminación precisos.

Es en ese sentido que el presente proyecto propone estudiar diferentes estrategias de transformación de datos y modelos de aprendizaje de máquina que puedan estimar el nivel de contaminación por $PM_{2,5}$, que es el más común y peligroso de todos los contaminantes, en zonas urbanas no censadas. Para ello, se entrenará y validará los modelos con un dataset de mediciones de calidad del aire recolectado en la ciudad de Beijing, China[1], la cual contiene información de concentraciones de contaminantes y variables meteorológicas con una estructura de datos bastante similar a la que se tendrá una vez que la red medición de Lima sea establecida. Cabe resaltar que como trabajo futuro, se planea usar los modelos probados con el dataset de Beijing, China en Lima, Perú.

Los potenciales beneficiarios de tener modelos de predicción espacial precisos dentro del sistema de monitoreo de calidad del aire serían los organismos de vigilancia, supervisión, fiscalización y control ambiental, tales como el SENAMHI, Ministerio del Ambiente y el Organismo de Evaluación y Fiscalización Ambiental (OEFA). Otra aplicación importante sería en el área de salud pública, donde el sistema podría ser usado para lanzar alertas a los ciudadanos sobre los lugares y horas donde los límites permitidos están siendo sobrepasados. Los beneficiarios de ello serían las autoridades de salud y los grupos sensibles a la contaminación del aire. Una tercera aplicación sería en el planeamiento urbano, donde la información histórica capturada puede ayudar a los planificadores a identificar zonas frecuentemente contaminadas y crear acciones paliativas.

2. Trabajos relacionados

En esta sección, se describirá brevemente los artículos más relevantes de los últimos 3 años, así como los hallazgos más importantes para la investigación.

Liu, B. C. et al. [3] propone un modelo que usa el algoritmo Support Vector Regression (SVR) para predecir el índice de calidad del aire (AQI) en tres ciudades de China.

[1] ver https://biendata.com/competition/kdd_2018/

En los datasets, se tomó en cuenta información de calidad de aire de las tres ciudades capturada cada hora, y condiciones climáticas como datos de entrada. Se experimentó con cuatro datasets por ciudad, por ejemplo, para predecir el AQI de la ciudad A, se usó los datasets de las ciudades A, A+B, A+C, A+B+C, donde B y C son ciudades cercanas. Se concluyó que para determinar el AQI de la ciudad que tiene un mayor nivel de contaminación es mejor no tomar en cuenta datasets que tienen menor nivel de contaminación o se encuentran muy alejadas.

Li, X. et al. [4] propone una versión extendida del modelo Long Short Term Memory (LSTME), el cual incluye las correlaciones espacio-temporales para lograr una mejor predicción del nivel de concentración de $PM_{2,5}$ en China. En su dataset, se tomó en cuenta el nivel de concentración de $PM_{2,5}$ capturada cada hora, y factores meteorológicos de Beijing, China. Se usó la correlación de Pearson para determinar la correlación espacial de las concentraciones de $PM_{2,5}$ en las 12 estaciones, y se determinó que había una correlación fuerte, lo cual indicaba que solo era necesario un modelo para todas las estaciones, y no un modelo para cada estación. Por otro lado, se usó funciones de autocorrelación para medir la correlación temporal entre las series temporales de concentración de $PM_{2,5}$ en cada estación, y se determinó que mientras mayor sea el lag[2], menor influencia tenía en el estado actual. Se usó capas Long Short Term memory (LSTM) para extraer características representativas de la data histórica, luego se agregó los factores meteorológicos y se usó capas densas para obtener representaciones de todas las características combinadas, y finalmente, se usó una capa densa para obtener el dato de salida.

Soh, P. W. et al. [9] propone usar una combinación de redes neuronales para predecir hasta 48 horas después el nivel de concentración de $PM_{2,5}$ en China. Su dataset incluye nivel de concentración de $PM_{2,5}$ y PM_{10} capturada cada hora, y data meteorológica de Taiwan y Beijing, así como información relacionada a la elevación del terreno con el fin de determinar su nivel de impacto en la calidad del aire. El modelo propuesto esta conformado por Artifical Neural Network (ANN), Convolutional Neural Network (CNN) y Long Short Term Memory (LSTM). Primero, se determinó las k estaciones más relacionadas a la estación objetivo usando k-Nearest Neighbor por distancia euclidiana (kNN-ED) para terrenos planos, y k-Nearest Neighbor por distancia DTW [12, 13] (kNN-DTWD) para terrenos complejos. Una vez seleccionados dichas estaciones, se utilizó capas ANN para extraer las características representativas de calidad del aire de las estaciones relacionadas a la estación objetivo. Así también, se usó las capas LSTM para extraer características representativas de la data histórica de la estación objetivo. Por otro lado, se extrajo información del terreno a los alrededores de la estación objetivo, y se usó capas CNN para extraer características relacionadas a la

[2]Lag es expresado en unidades de tiempo (ejm: horas) y corresponde a la cantidad de data histórica que permitimos usar al modelo para la predicción

interacción entre el terreno y la calidad del aire. Finalmente, teniendo como datos de entrada las características obtenidas anteriormente, se usó capas densas para obtener el dato de salida. Estas capas finales aprenden los pesos de la data de entrenamiento, en ocasiones tiene más peso la data espacial que temporal (por ejemplo: días con viento) o viceversa. Según el autor, este modelo propuesto ST-DNN es aplicable también a otros contaminantes.

Wang, J. et al. [10] propone el modelo de aprendizaje profundo espacio-temporal ensamblado (STE) para predecir hasta 48 horas después el nivel de concentración de $PM_{2,5}$ en China. Su dataset incluye nivel de concentración de $PM_{2,5}$, PM_{10}, SO_2, CO, NO_2, y O_3 capturada cada hora, y data meteorológica capturada cada 3 horas. El modelo propuesto STE consiste en tres partes. Primero, se usó un método de ensamblado por patrones climatológicos, es decir, se clasificó la data histórica de calidad del aire en diferentes sub-datasets teniendo como base el patrón climatológico. El patrón de cada dato está representado por la causalidad de Granger obtenida a partir del factor climatológico y calidad del aire en un determinado periodo. Para que haya suficientes datos para cada sub-dataset, se uso el algoritmo de agrupación Fuzzy C-Means. Como segundo componente, se usó una vez más la causalidad de Granger para determinar las estaciones y áreas más correlacionadas con la estación objetivo. Como tercer componente, se usó capas LSTM para aprender las dependencias a corto y largo plazo de la calidad del aire. Una vez entrenado los modelos para cada sub-dataset, se usó Support Vector Regresion (SVR) para agregar todos los resultados de todos los modelos dinámicamente y obtener el dato de salida final.

En el cuadro 1, se puede ver información importante y resumida que se pudo obtener de todos los artículos revisados. Se observa que los artículos más relevantes son de China, pues tienen problemas graves de contaminación de aire, en especial con el contaminante $PM_{2,5}$. Estos artículos se enfocan tanto en la predicción espacial como temporal, dando mucha mayor importancia a lo segundo. Han habido pocos esfuerzos en cuanto a la estimación espacial probablemente debido a que cuentan con una gran cantidad de estaciones de calidad del aire[3]. Este no es el caso de Perú, donde contamos con muy pocas estaciones, en ocasiones no funcionales[4]. Por otro lado, entre las variables meteorológicas que usaron para los estudios se encuentran el clima, temperatura, humedad, elevación de terreno y velocidad y dirección del viento, entre otros.

[3]ver http://aqicn.org/map/china/
[4]ver http://aqicn.org/map/peru/

Autor	Año	Agentes ambientales	Data meteorológica y otros	Análisis espacial	Análisis temporal	Modelo	Ciudad	Predicción
Liu, B. C. et al. [3]	2017	$PM_{2,5}$, PM_{10}, SO_2, CO, NO_2, y O_3	Temperatura mínima, temperatura máxima, clima, dirección del viento, y poder del viento	-	-	SVR	Beijing, Tianjin y Shijiazhuang	AQI
Li, X. et al. [4]	2017	$PM_{2,5}$	Temperatura, humedad, velocidad del viento y visibilidad	Correlación de Pearson	Autocorrelación	LSTM + FC	Beijing (12 estaciones)	$PM_{2,5}$
Xu, Y. et al. [5]	2017	$PM_{2,5}$, PM_{10}, SO_2, CO, NO_2, y O_3	-	-	-	ICEEMD + SVM + WOA	Taiyuan, Harbin y Chongqing	$PM_{2,5}$
Soh, P. W. et al. [9]	2018	$PM_{2,5}$, PM_{10}	Temperatura, humedad, velocidad del viento, dirección del viento, y elevación del terreno	kNN por distancia euclidiana	kNN por distancia DTW	ANN + LSTM + CNN + FC	Taiwan y Beijing	$PM_{2,5}$
Wang, J. et al. [10]	2018	$PM_{2,5}$, PM_{10}, SO_2, CO, NO_2, y O_3	Temperatura, humedad, velocidad del viento, dirección del viento	Causalidad de Granger	LSTM	Ensemble by weather patterns + Granger causality + LSTM + SVR	Beijing (35 estaciones)	$PM_{2,5}$
Wen, C. et al. [11]	2019	$PM_{2,5}$	Temperatura, humedad, velocidad del viento, altura de la capa límite planetaria, y profundidad óptica de aerosoles	Distancia euclidiana y Correlación de Pearson	LSTM	CNN + LSTM + ANN	Beijing (12 estaciones)	$PM_{2,5}$

Cuadro 1: Artículos relacionados a la predicción de nivel de contaminación de aire de los últimos 3 años

3. Metodología

Para obtener el modelo con mejor capacidad de estimación espacial, se evaluó exhaustivamente 3 factores que consideramos importantes. Un factor es el uso de las variables urbanas, otro es la normalización por IDW que se describirá luego, y el otro es la cantidad de estaciones más cercanas que se usará para el entrenamiento y validación del modelo, dicha cantidad será denominada k en lo sucesivo.

Se usó el conjunto de datos dado en el concurso "KDD CUP of Fresh Air" de China[5]. Este cuenta con 2 subconjuntos de datos, uno contiene concentraciones de los diversos contaminantes del aire, tales como $PM_{2,5}$, PM_{10}, SO_2, CO, NO_2, O_3, y el otro contiene variables urbanas que podrían interferir en el nivel de contaminación, tales como clima, temperatura, humedad, velocidad y dirección del viento.

Nos enfocaremos en estimar el nivel de $PM_{2,5}$. Para ello, se implementó un flujo de pre-procesamiento de datos, que se encargó de separar los datos relacionados a $PM_{2,5}$ y después concatenar dicho dataset de contaminantes con el de variables urbanas, así como identificar, modificar o eliminar registros faltantes y corrompidos.

Luego, se implementó la normalización por IDW, adaptación del modelo *Inverse Distance Weighting (IDW)*[15], que consiste en transformar cada dato del dataset usando la siguiente formula:

$$c_p = \frac{\frac{c_i}{d_i}}{\sum_{i=1}^{k} \frac{1}{d_i}} \tag{1}$$

Donde c_p es el dato transformado, c_i es el contaminante $PM_{2,5}$ o variable urbana de una estación i, k es la cantidad de estaciones más cercanas al punto de predicción p, y d_i es la distancia de la estación i al punto de predicción p donde se quiere estimar el contaminante.

Usando la normalización por IDW, diferentes valores de k y la inclusión de variables urbanas, se creó múltiples datasets a partir del dataset de contaminantes y variables meteorológicas pre-procesado con la finalidad de determinar la relevancia de estos 3 factores. Estos datasets pueden agruparse en los siguientes:

- Dataset de contaminantes $PM_{2,5}$
- Dataset de contaminantes $PM_{2,5}$ normalizado por IDW
- Dataset de contaminantes $PM_{2,5}$ y variables urbanas
- Dataset de contaminantes $PM_{2,5}$ normalizado por IDW y variables urbanas
- Dataset de contaminantes $PM_{2,5}$ y variables urbanas normalizados por IDW
- Dataset de contaminantes $PM_{2,5}$ y variables urbanas (de la estación objetivo)
- Dataset de contaminantes $PM_{2,5}$ normalizado por IDW y variables urbanas (de

[5]ver https://biendata.com/competition/kdd_2018/

la estación objetivo)

- Dataset de contaminantes $PM_{2,5}$ y dirección del viento
- Dataset de contaminantes $PM_{2,5}$ normalizado por IDW y dirección del viento

Entre los modelos que se estudiaron se encuentran los de la línea base tales como *Nearest Neighbor (k-NN)* e *Inverse Distance Weighting (IDW)*; los modelos de aprendizaje de máquina tales como *Linear Regression (LR)*, *Support Vector Regression (SVR)*, *Random Forest (RF)*, *Xtreme Gradient Boosting (XGBoost)* y *Light Gradient Boosting Machine (LightGBM)*; y finalmente un modelo de red neuronal de tipo *Feed-Forward (FF-NN)*.

Para la evaluación de dichos modelos, se usó la métrica R^2 (coeficiente de determinación)[6]. Se desarrolló un pipeline de entrenamiento y validación que usó las diversas versiones de datasets creados previamente, y siguió una estrategia que consiste en tener un modelo por cada k y estación de la red como punto de predicción, y luego calcular el promedio de los R^2 obtenidos en las diferentes estaciones para cada k. Es decir, cada modelo estimó el nivel de $PM_{2,5}$ en las coordenadas de cada una de las estaciones de la red. Para ello, se entrenó el modelo con datos de las k estaciones más cercanas, y luego se validó con los datos de la estación objetivo. Una vez determinado el R^2 de todas las estaciones, se calculó el promedio. Este proceso se repitió para todos los valores de k posibles.

Una vez determinado el R^2 promedio por cada k, se procedió a realizar un análisis del comportamiento y el impacto del valor de k y de las variables urbanas en la predicción del nivel de $PM_{2,5}$ para los distintos modelos y datasets.

4. Experimentación

En esta sección, se describirá a detalle los diversos enfoques que se tuvieron en cuenta para obtener el modelo con mejor capacidad de estimación espacial dado el uso de variables urbanas, la normalización por IDW y un determinado k. Cabe resaltar que solo se mencionarán los experimentos que dieron mejores resultados.

4.1. Línea base

Primero, se probó la estrategia k-NN (o vecino más cercano), la forma más simple de interpolación espacial. Consiste en que el punto objetivo donde se quiere estimar el nivel de contaminación toma el valor de la suma promedio del nivel de contaminación de las k estaciones más cercanas. La fórmula sería la siguiente:

[6]Estadístico que determina la calidad del modelo para replicar los resultados, y la proporción de variación de los resultados que puede explicarse por el modelo [14]

$$c_p = \frac{\sum_{i=1}^{k} c_i}{k} \tag{2}$$

Donde c_p es el contaminante $PM_{2,5}$ en el punto objetivo, c_i es el contaminante en una estación i, y k es la cantidad de estaciones más cercanas.

Luego, se probó el método determinístico *Inverse Distance Weighting (IDW)* [15]. Consiste en que el punto objetivo donde se quiere estimar el nivel de contaminación toma el valor obtenido por la siguiente fórmula:

$$c_p = \frac{\sum_{i=1}^{k} \frac{c_i}{d_i}}{\sum_{i=1}^{k} \frac{1}{d_i}} \tag{3}$$

Donde c_p es el contaminante $PM_{2,5}$ en el punto objetivo, c_i es el contaminante en una estación i, k es la cantidad de estaciones más cercanas, y d_i es la distancia de dicha estación i al punto donde se quiere estimar el contaminante.

4.2. Modelos de aprendizaje de máquina

Se definió una estrategia que consiste en estimar el nivel de $PM_{2,5}$ en la posición de una de las estaciones. El procedimiento consiste en entrenar un modelo con las k estaciones más cercanas a la del objetivo, y luego probar dicho modelo para estimar el nivel de $PM_{2,5}$ en la posición de la estación objetivo. En la Figura 1, se pueda visualizar a grandes rasgos el procedimiento teniendo como objetivo estimar el nivel de contaminación en la posición de la estación 12, dado $k = 11$.

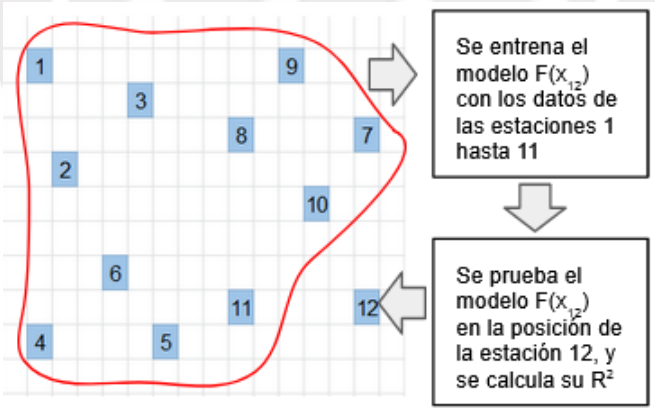

Figura 1: Predicción de nivel de $PM_{2,5}$ en la posición de la estación 12, dado $k = 11$

Para analizar el impacto de k, la normalización por IDW y las variables urbanas, se repitió el procedimiento antes descrito con todos los valores posibles de k y diversas versiones de datasets creados. A continuación, se muestra los algoritmos empleados para

el procesamiento de datos, así como el entrenamiento y validación de los modelos.

4.2.1. Procesamiento de datos

En el Algoritmo 1 se puede apreciar el procesamiento de datos general que se usó para la creación de los datasets de entrenamiento y de validación.

Algoritmo 1: Procesamiento de datos

```
   /* Por cada valor de cantidad de estaciones más cercanas a la
      estación objetivo                                          */
1  for k do
   |  /* Por cada estación objetivo                              */
2  |  for t-estacion do
3  |  |  Remover la estación objetivo del dataset;
   |  |  /* Por cada estación restante                          */
4  |  |  for r-estacion do
5  |  |  |  Crear subdataset (k, t-estacion, r-estacion)
6  |  |  end
7  |  |  Concatenar los subdatasets para crear el dataset de entrenamiento;
8  |  |  Crear el dataset de prueba con la misma estructura, a partir de la latitud y
   |  |   longitud de la estación objetivo;
9  |  end
10 end
```

El algoritmo anterior fue adaptado para diferentes estrategias. Estas adaptaciones radican en las variantes de la función de creación de subdatasets, que consisten en diversas formas de ordenamiento de las estaciones, sin tener en cuenta $r - estacion$ y $t - estacion$.

Una de las variantes consiste en ordenar dichas estaciones por distancia euclidiana, es decir, el $PM_{2,5}$ y las variables urbanas de la estación más cercana a $r - estacion$ serán los primeros atributos del dataset, y los de la estación más lejana serán los últimos.

Algoritmo 2: Ordenamiento por distancia euclidiana

1 Calcular la distancia entre r-estacion y las demás k estaciones, sin incluir t-estacion;
2 Crear un dataset con la estructura: st-1-$PM_{2,5}$, st-1-var, ..., st-k-$PM_{2,5}$, st-k-var;
  donde se representa el nivel de $PM_{2,5}$ y variables urbanas de las estaciones ordenadas del más cercano al más lejano de r-estacion;

Otra variante, que es muy parecida a la anterior, consiste en ordenar las estaciones por distancia euclidiana, es decir, el $PM_{2,5}$ de la estación más cercana a $r - estacion$

será el primer atributo del dataset, y el de la estación más lejana será el último. En cuanto a las variables urbanas, se tomará solo las de $r - estacion$.

Algoritmo 3: Ordenamiento por distancia euclidiana, considerando solo variables urbanas de r-estacion

1 Calcular la distancia entre r-estación y las demás k estaciones, sin incluir t-estacion;
2 Crear un dataset con la estructura: st-1-$PM_{2,5}$, ..., st-k-$PM_{2,5}$; donde se representa el nivel de $PM_{2,5}$ de las estaciones ordenadas del más cercano al más lejano de r-estacion;
3 Agregar al dataset anterior: st-r-var; donde se representa las variables urbanas de r-estacion;

Finalmente, esta última variante consiste en ordenar las estaciones por distancia euclidiana teniendo en cuenta la dirección del viento de $r - estacion$, es decir, al igual que los casos anteriores, las estaciones serán ordenadas del más cercano al más lejano a $r - estacion$, pero además, aquellas estaciones que no se encuentren favorecidas por la dirección del viento se les multiplicará la distancia con un factor de 1.5, de tal forma que tengan un menor impacto al momento de estimar el $PM_{2,5}$ sobre la estación objetivo. Por ejemplo, si la dirección del viento va de Oeste a Este, las estaciones que se encuentren en el lado Este de $r - estacion$ serían las no favorecidas por el viento, y por ende su distancia a $r - estacion$ sería multiplicada por 1.5 ya que menos relevantes que las estaciones ubicadas en el lado Oeste.

Algoritmo 4: Ordenamiento por distancia euclidiana y dirección del viento

1 Calcular la distancia entre r-estación y las demás k estaciones, sin incluir t-estacion;
2 **if** *k-estacion no es favorecida por el viento* **then**
3 | distancia = distancia * 1.5;
4 **end**
5 Crear un dataset con la estructura: st-1-$PM_{2,5}$, ..., st-k-$PM_{2,5}$; donde se representa el nivel de $PM_{2,5}$ de las estaciones ordenadas del más cercano al más lejano de r-estacion, teniendo en cuenta la dirección del viento en r-estacion;

4.2.2. Entrenamiento y prueba de modelos

En el Algoritmo 5 se puede apreciar el proceso para la fase de entrenamiento y validación de los modelos, así como el cálculo del R^2 promedio.

Algoritmo 5: Entrenamiento y prueba de modelos

```
   /* Por cada valor de cantidad de estaciones más cercanas a la
      estación objetivo                                          */
1  for k do
       /* Por cada estación objetivo                             */
2      for t-estación do
3          Entrenar modelo con el dataset de entrenamiento (k, t-estacion);
4          Validar modelo con el dataset de prueba (k, t-estacion);
5          Calcular la métrica R^2;
6      end
7      Calcular el promedio de las métricas R^2;
8  end
```

5. Resultados y Discusiones

En esta sección, se mostrará los mejores resultados obtenidos de los experimentos realizados.

En la Figura 2, se puede ver el comportamiento del R^2 promedio en función de k para los modelos *k-NN*, *IDW* y *Linear Regression* usando el dataset de contaminantes normalizado por IDW. Estos fueron los que presentaron el mayor valor de R^2 en un determinado k, aunque cabe resaltar que a medida que aumenta el k, el R^2 decae notablemente.

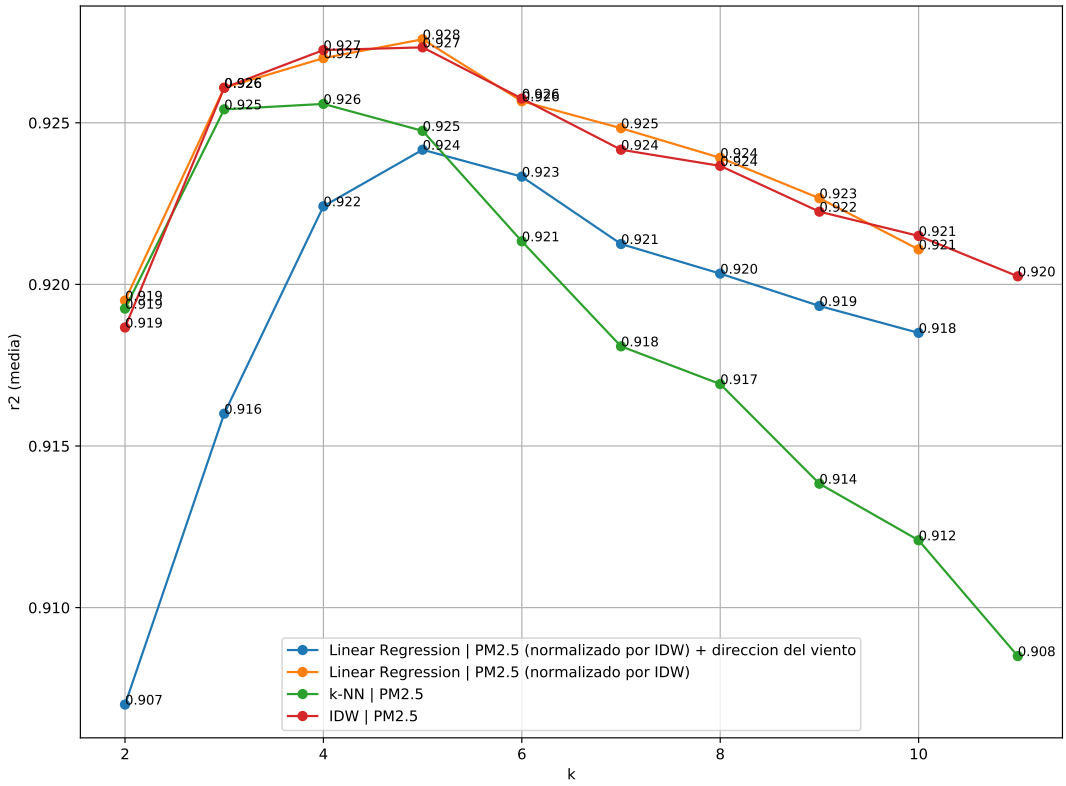

Figura 2: R^2 promedio por k, de los modelos k-NN, IDW y Linear Regression

Por otro lado, en la Figura 3, se ve una comparación del desempeño del mejor modelo y el modelo *LightGBM* con los diferentes datasets creados. Se puede ver que el valor del R^2 promedio, a pesar de que sea menor, no decae notablemente a medida que sube el k en comparación de *Linear Regression*.

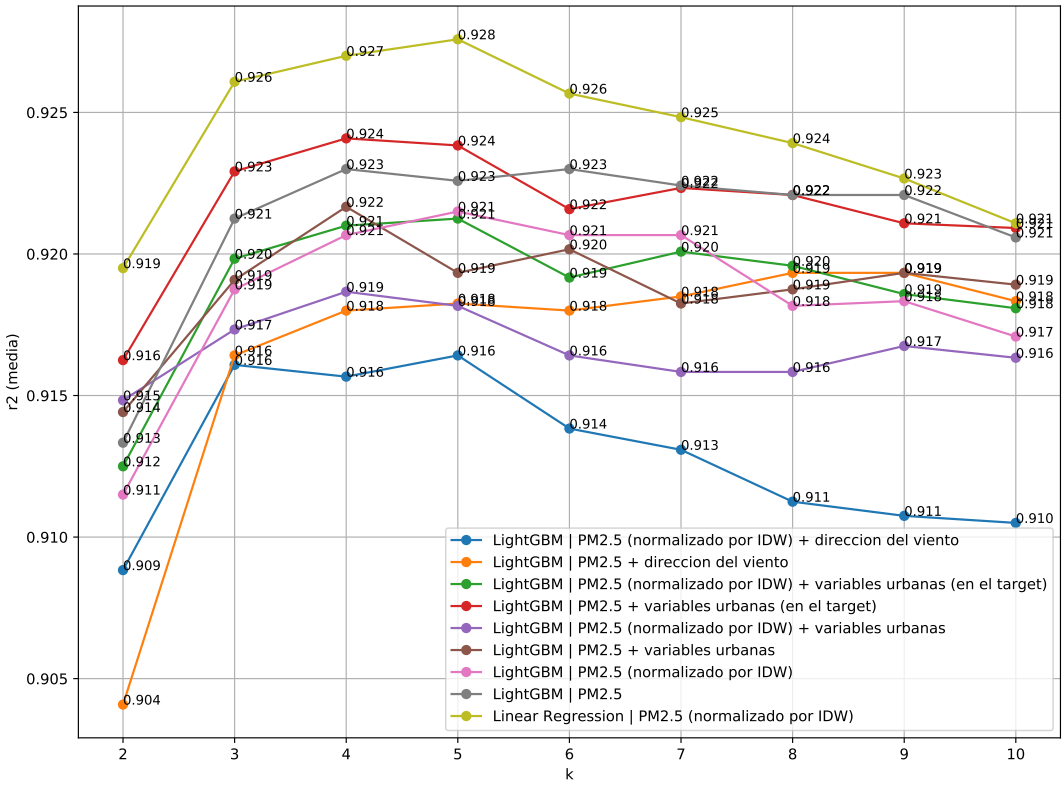

Figura 3: R^2 promedio por k, del modelo LightGBM

En la Figura 4, también se ve una comparación del desempeño del mejor modelo y el modelo *XGBoost* con los diferentes datasets creados. Se puede ver que el valor del R^2 promedio es mayor al de *LightGBM* pero menor al de *Linear Regression*, y al igual que el caso anterior, no decae notablemente a medida que sube el k.

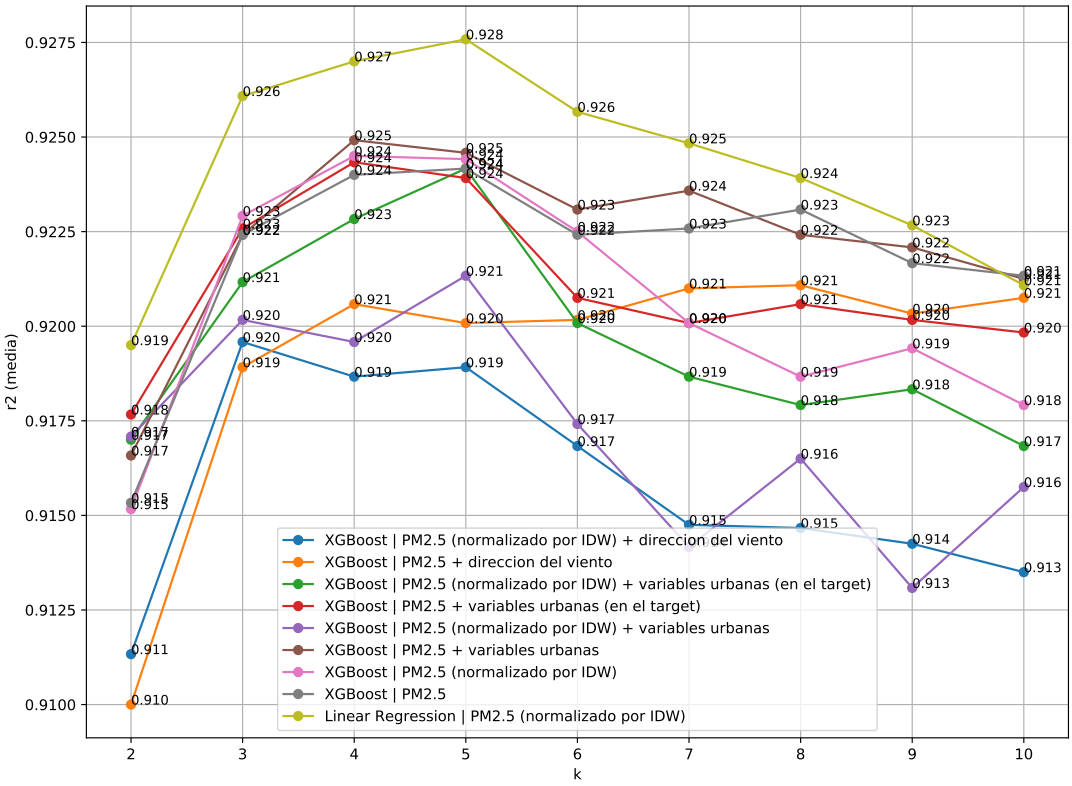

Figura 4: R^2 promedio por k, del modelo XGBoost

Se logró superar la línea base usando diferentes enfoques, ya sea por mejor R^2 en un determinado k, o por ser modelos robustos a la variabilidad de k. Además, se ve que en todos los experimentos, el R^2 es bastante bajo cuando k es igual a 1 o 2, a partir de 3 se empieza a ver mejora en los resultados. A continuación, en el Cuadro 2, se muestra una tabla resumen de todos los experimentos realizados, donde se verá el mejor k, el R^2 promedio y observaciones acerca del comportamiento del modelo dado un k. Específicamente se tendrá 3 tipos de observaciones, cuando R^2 es estable quiere decir que la variación de R^2 a medida que k aumenta es casi nula, cuando R^2 es casi estable quiere decir que la variación de R^2 a medida que k aumenta es baja pero no pasa desapercibido, para los demás casos se escribirá explícitamente el comentario.

Modelo	Mejor k	R^2	Observaciones
Línea base			
k-NN	4	0.926	A mayor k, mucho menor R^2.
IDW	4,5	0.927	A mayor k, menor R^2.
Modelos con $PM_{2,5}$			
LightGBM	4,5,6,7	0.923	R^2 estable.
XGBoost	4,5	0.924	R^2 estable.
Modelos con $PM_{2,5}$ normalizado por IDW			
LR	5	0.928	A mayor k, menor R^2.
LightGBM	4,5,6,7	0.921	R^2 casi estable.
XGBoost	4,5	0.924	R^2 casi estable.
FF-NN	4,5	0.927	A mayor k, menor R^2.
Modelos con $PM_{2,5}$ y variables urbanas			
LightGBM	4	0.922	$R^2 = 0{,}919$ estable a partir de $k = 5$ en adelante.
XGBoost	4,5	0.925	R^2 casi estable.
Modelos con $PM_{2,5}$ normalizado por IDW y variables urbanas			
LR	3	0.921	A mayor k, mucho menor R^2.
LightGBM	4	0.919	$R^2 = 0{,}916$ estable a partir de $k = 6$ en adelante.
XGBoost	5	0.925	$R^2 = 0{,}917$ estable a partir de $k = 7$ en adelante.
Modelos con $PM_{2,5}$ y variables urbanas normalizados por IDW			
LR	3	0.921	A mayor k, mucho menor R^2.
LightGBM	5	0.918	R^2 casi estable.
XGBoost	5	0.923	R^2 casi estable.
Modelos con $PM_{2,5}$ y variables urbanas (de la estación objetivo)			
LightGBM	4,5	0.924	$R^2 = 0{,}922$ estable a partir de $k = 6$ en adelante.
XGBoost	4,5	0.924	$R^2 = 0{,}920$ estable a partir de $k = 6$ en adelante.
Modelos con $PM_{2,5}$ normalizado por IDW y variables urbanas (de la estación objetivo)			
LightGBM	4,5	0.921	R^2 casi estable.
XGBoost	5	0.924	A mayor k, menor R^2.
Modelos con $PM_{2,5}$ y direccion del viento			
LightGBM	8,9	0.919	R^2 casi estable.
XGBoost	4,7,8,10	0.921	R^2 estable.
Modelos con $PM_{2,5}$ normalizado por IDW y direccion del viento			
LR	5	0.924	A mayor k, menor R^2.
LightGBM	3,4,5	0.916	A mayor k, menor R^2.
XGBoost	3	0.920	A mayor k, menor R^2.

Cuadro 2: Mejor k y R^2 por modelo

6. Conclusiones

Teniendo como referencia los gráficos y la tabla resumen de la sección anterior, se puede concluir lo siguiente:

1. El modelo *Linear Regression* usando solo el dataset de contaminante $PM_{2,5}$ nor-

malizado por IDW presenta el mejor desempeño predictivo entre todos los modelos probados para un $k = 5$. En segundo lugar quedaría el modelo línea base *IDW* y el modelo de red neuronal *FF-NN*.

2. El agregar las variables urbanas causan un efecto negativo al modelo *Linear Regression*, pues a medida que aumenta el k, el R^2 decrece significativamente hasta llegar al punto que sea menor a cero.
3. Los modelos *XGBoost* y *LightGBM*, a pesar de tener un menor nivel de predictibilidad, son más robustos, pues se mantienen más estables a pesar que el k aumente.
4. Las variables urbanas aportaron un beneficio marginal al modelo *XGBoost*, mientras que en el caso de *LightGBM* dio un efecto negativo.
5. Normalizar por IDW la variable contaminante $PM_{2,5}$ causó un efecto negativo en los modelos *XGBoost* y *LightGBM*.
6. Usar la dirección del viento como factor para darle menos peso a las estaciones a favor del viento causó un efecto negativo en los modelos *Linear Regression*, *XGBoost* y *LightGBM*.

7. Trabajos futuros

Como trabajo futuro, se planea usar los modelos estudiados en esta investigación en las zonas urbanas de Lima, Perú.

Es decir, se tomará los datos almacenados en los módulos de medición de calidad del aire de Lima, que al igual que en el dataset de China, estaría conformado por las concentraciones de los diversos contaminantes del aire, tales como $PM_{2,5}$, PM_{10}, SO_2, CO, NO_2, O_3, y variables urbanas, tales como temperatura, humedad, velocidad y dirección del viento, y tráfico vehicular. Después, se procederá a adaptar el dataset de Lima, de tal forma que tenga una estructura igual al de China, con el fin de que sea posible emplear el pipeline de pre-procesamiento de datos.

Con el dataset preparado previamente, se re-entrenará y evaluará los mejores modelos y se hará un análisis comparativo. El modelo óptimo será aquel que tenga mayor R^2 (coeficiente de determinación) y mayor simplicidad estructural. Esto último es necesario ya que el costo computacional de inferencia sería menor, lo que conlleva a un menor tiempo de latencia en la actualización de los mapas de nivel de contaminación producidos.

Referencias

[1] Baumann, L. M., Robinson, C. L., Combe, J. M., Gomez, A., Romero, K., Gilman, R. H., ... & Barnes, K. (2011). Effects of distance from a heavily transited avenue on asthma and atopy in a periurban shantytown in Lima, Peru. Journal of Allergy and Clinical Immunology, 127(4), 875-882.

[2] Bellinger, C., Jabbar, M. S. M., Zaïane, O., & Osornio-Vargas, A. (2017). A systematic review of data mining and machine learning for air pollution epidemiology. BMC public health, 17(1), 907.

[3] Liu, B. C., Binaykia, A., Chang, P. C., Tiwari, M. K., & Tsao, C. C. (2017). Urban air quality forecasting based on multi-dimensional collaborative support vector regression (svr): A case study of beijing-tianjin-shijiazhuang. PloS one, 12(7), e0179763.

[4] Li, X., Peng, L., Yao, X., Cui, S., Hu, Y., You, C., & Chi, T. (2017). Long short-term memory neural network for air pollutant concentration predictions: Method development and evaluation. Environmental pollution, 231, 997-1004.

[5] Xu, Y., Yang, W., & Wang, J. (2017). Air quality early-warning system for cities in China. Atmospheric environment, 148, 239-257.

[6] Freeman, B. S., Taylor, G., Gharabaghi, B., & Thé, J. (2018). Forecasting air quality time series using deep learning. Journal of the Air & Waste Management Association, 1-21. 1982, p. 301].

[7] Reátegui-Romero, W., Sánchez-Ccoyllo, O. R., de Fatima Andrade, M., & Moya-Alvarez, A. (2018). PM2.5 Estimation with the WRF/Chem Model, Produced by Vehicular Flow in the Lima Metropolitan Area. Open Journal of Air Pollution, 7(03), 215.

[8] Sánchez-Ccoyllo, O. R., Ordoñez-Aquino, C. G., Muñoz, Á. G., Llacza, A., Andrade, M. F., Liu, Y., ... & Brasseur, G. (2018). Modeling Study of the Particulate Matter in Lima with the WRF-Chem Model: Case Study of April 2016. International Journal of Applied Engineering Research, 13(11), 10129-10141.

[9] Soh, P. W., Chang, J. W., & Huang, J. W. (2018). Adaptive deep learning-based air quality prediction model using the most relevant spatial-temporal relations. IEEE Access, 6, 38186-38199.

[10] Wang, J., & Song, G. (2018). A deep spatial-temporal ensemble model for air quality prediction. Neurocomputing, 314, 198-206.

[11] Wen, C., Liu, S., Yao, X., Peng, L., Li, X., Hu, Y., & Chi, T. (2019). A novel spatiotemporal convolutional long short-term neural network for air pollution prediction. Science of The Total Environment, 654, 1091-1099.

[12] Rakthanmanon, T., Campana, B., Mueen, A., Batista, G., Westover, B., Zhu, Q., ... & Keogh, E. (2012, August). Searching and mining trillions of time series subsequences under dynamic time warping. In Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 262-270). ACM.

[13] Keogh, E., & Ratanamahatana, C. A. (2005). Exact indexing of dynamic time warping. Knowledge and information systems, 7(3), 358-386.

[14] Steel, R. G. D., & Torrie, J. H. (1960). Principles and procedures of statistics. Principles and procedures of statistics.

[15] Shepard, D. (1968, January). A two-dimensional interpolation function for irregularly-spaced data. In Proceedings of the 1968 23rd ACM national conference (pp. 517-524). ACM.

[16] OMS. (2018). Nueve de cada diez personas de todo el mundo respiran aire contaminado. Recuperado de: https://www.who.int/es/news-room/detail/02-05-2018-9-out-of-10-people-worldwide-breathe-polluted-air-but-more-countries-are-taking-action.

[17] Naciones Unidas. (2018). La Agenda 2030 y los Objetivos de Desarrollo Sostenible: una oportunidad para América Latina y el Caribe(LC/G.2681-P/Rev.3), Santiago.