Causalidad y prediccion de CO, NO_2 y PM_{10} en la Ciudad de Buenos Aires

Análisis de la calidad de aire mediante Machine Learning

Ezequiel Vannucchi
Sr. Growth Data Analyst at Glovo
Ing. Industrial UTN BA

Ivan Ariel Bergera Vila
Analista Comercial at Telefónica
Ing. Industrial UTN BA

Aldana Cavallucci
Pasante de desarrollo organizacional at YPF
Ing. Industrial UTN BA

Abstract
El objetivo de este proyecto es identificar cuales son las variables, o combinación de ellas, que mejor explican los principales contaminantes de la calidad del aire de la ciudad de Buenos Aires y evaluar la factibilidad de un modelo que permita la predicción de estos sin la necesidad de monitoreos constantes.

1 INTRODUCCIÓN

Dada las problemáticas referidas a la calidad del aire que se respira en las ciudades, así como las posibles consecuencias que esto puede acarrear para las personas, y considerando la diversa cantidad de sustancias gaseosas como particulados que se pueden encontrar en el mismo, el objetivo de este estudio es analizar si a partir de variables sencillamente medibles en forma diaria poder realizar una predicción de los niveles de monóxido de carbono (CO), dióxido de nitrógeno (NO_2) y material particulado menor a 10 micrómetros (PM_{10}).

Ha habido antecedentes de este tipo de desarrollo en España para la predicción de NO_2[1] pero no se encontraron registros de

De poder lograr esto el objetivo sería poder extrapolar el modelo a otros puntos de la ciudad para poder lograr mapeo que permita identificar las zonas de mayor riesgo de cada material y poder así desarrollar planes de acción en base a cada necesidad.

2 LA DATA

2.1 Fuentes

Los datasets utilizados provienen de las mediciones de calidad de aire para el 2019 y Conteo Vehicular 2019 (provistos por el gobierno de la Ciudad de Buenos Aires) y las mediciones de los factores meteorológicos (provistos por el Servicio Meteorológico Nacional).

Si bien el dataset inicial (Calidad de aire) contaba con mediciones en otros puntos de la ciudad esta información no era compatible con la provista por los otros datasets. Por esto mismo se decidió trabajar con la estación de medición ubicada en La Boca.

De esta forma se concluyó con un dataset crudo con 2868 muestras con las mediciones correspondientes a los contaminantes previamente mencionados, variables temporales (día y hora), meteorológicas (Temperatura, humedad, presión atmosférica y dirección y fuerza del viento) y de mediciones del flujo vehicular en la zona (en cantidad de vehículos que circulan tanto en ambos sentidos de la autopista que pasa en las proximidades de la estación de medición).

Cabe aclarar que, si bien se consideró la utilización de otras variables, tales como m^2 de espacios verdes cercanos, presencia de industrias, habitantes de la zona y altitud entre otras, al quedar forzado este estudio a una única ubicación por las razones previamente explicadas, ninguna de estas variables hubiera presentado variación en el periodo de análisis. No obstante, en caso de escalar el proyecto, estas serían consideradas[1].

El detalle de como se realizó la combinación de estos 3 sets de datos puede encontrarse en un repositorio de GitHub al final del trabajo, así como los datasets originales[2].

2.2 Análisis Exploratorio de Datos (EDA)

La primera decisión que se tomo fue analizar la cantidad de valores faltantes (NaNs) de las distintas variables. De esta forma el de muestras se redujo a 2325 (19%) antes de arrancar con el análisis.

Fue en este punto que surgió la idea de conservar las muestras de nuestras variables que no tenían una medición que les correspondiera (es decir no había una etiqueta para ese set de features), y utilizar estas muestras para probar el modelo en funcionamiento una vez estuviera desarrollado.

Lo siguiente que se llevó a cabo fue un análisis par a par de cada una de nuestras mediciones para buscar tanto correlaciones como identificar la posible existencia de valores anómalos (outliers) dentro de las mediciones. Como puede verse en la **Figura 1** se pueden identificar unos pocos outliers tanto en las mediciones de CO como de NO_2, no así en lo que es PM_{10}, cuya distribución es más homogénea.

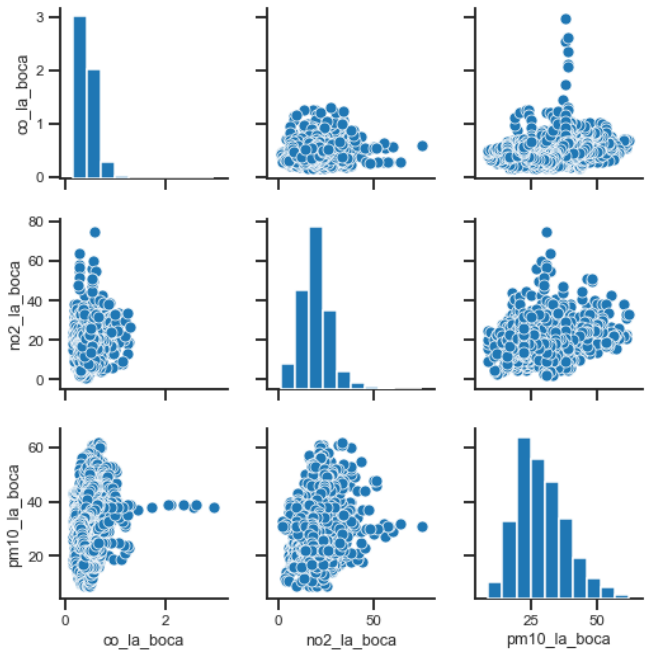

Figura 1

Luego se procedió a identificar como se ubicaba esta distribución en los distintos horarios durante la semana (. Para esto se utilizaron los mapas de calor que se ven a continuación (**Figura 2**).

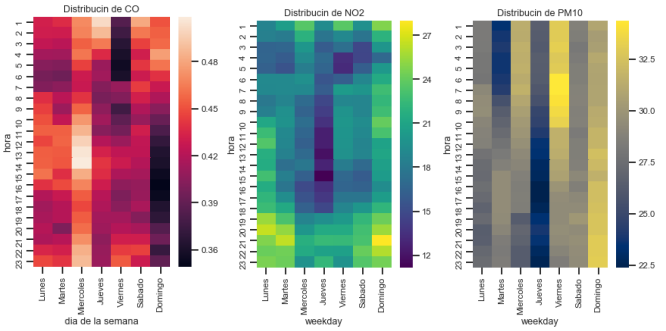

Figura 2

De esta forma se pudo identificar un comportamiento casi inverso entre el CO y el NO_2, el primero cuyos valores se reducen durante los fines de semana mientras que los del segundo aumentan. Por otro lado, la presencia de material particulado parece muy ligado a la presencia de NO_2 como ya se veía levemente en el gráfico de dispersión par-a-par anterior, compartiendo horarios de alta y baja presencia.

Finalmente, para terminar de comprender las mediciones y la relación de estas con las otras variables se procedió a realizar una matriz de correlación (**Figura 3**) entre las distintas variables. La misma se adjunta a continuación.

Es interesante enfocarnos en la correlación negativa que presenta en CO con respecto a la presión atmosférica. Esta variable tendrá mayor influencia a la hora de realizar el modelo de este contaminante. Si bien hay presencia de correlaciones más elevadas (los flujos vehiculares en ambas direcciones, y la temperatura y la presión en forma negativa), estas hasta cierto punto son esperables.

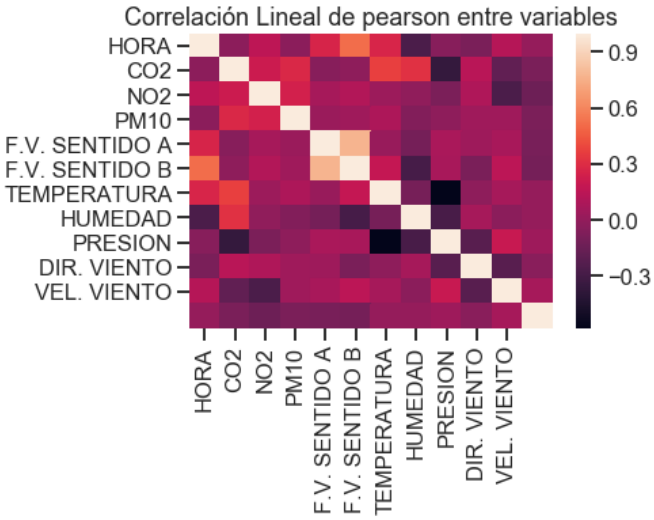

Figura 3

2.3 Análisis de Componentes Primarios (PCA)

Después de haber analizado la existencia o no de una correlación lineal entre las variables consideradas una a una se procedió a utilizar el método de PCA[2] para identificar si existen grupos o clusters de muestras.

Este método consiste en realizar combinaciones lineales de las variables para generar un numero reducido de variables que contengan el mayor porcentaje de la variabilidad total del dataset.

Para esto, así como para el resto de los análisis que se realizarán, se llevara a cabo una normalización de los valores, de forma que la variabilidad no quede determinada por unas pocas variables cuya dispersión numérica sea mayor, (otorgándoles más peso) sino que todas serán llevadas a valores entre 0 y 1 de forma que su variabilidad quede expresada en este intervalo. Esto se realiza de la siguiente forma

$$x_i' = \frac{(x_i - x_{min})}{(x_{max} - x_{min})}$$

De esta forma, y una vez aplicado el PCA se puede analizar la variabilidad del set de datos a partir de las nuevas variables generadas (primary components). Esto puede verse n la **Figura 4.**

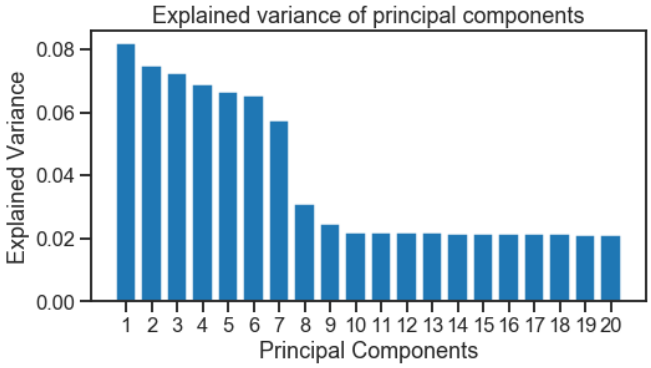

Figura 4

Graficando las 3 componentes principales en un gráfico de dispersión (**Figura 5**) se observa que las muestras se agrupan generando distintos clústers.

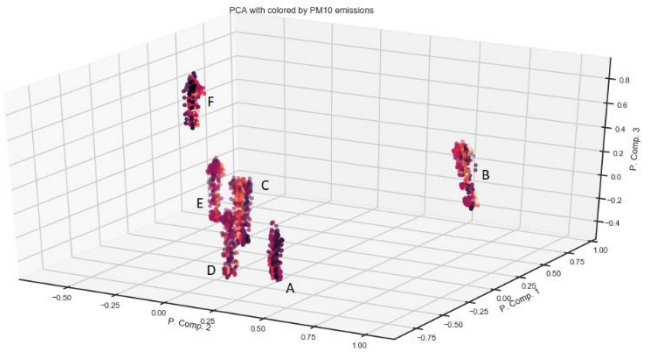

Figura 5

Mediante la coloración de las muestras se procedió a intentar identificar si estos clústeres agrupaban los valores de las distintas emisiones analizadas. Los mejores resultados los dio el análisis de las PM_{10} que son las que ese ve en la figura. Los valores negativos pueden verse agrupados en el clúster A mientras que los más positivos pueden identificarse tanto en el clúster B como en el C en menor medida. Los clusters E y D presentan valores más neutrales. Por ultimo el clúster F presenta una interesante combinación entre valores extremos tanto positivos como negativos.

3 EL MODELO

Se consideraron 2 metodologías a la hora de encarar el desarrollo del modelo.

La primera consideraba en establecer franjas de alta, media o bajas emisiones para cada componente e intentar predecir a que categoría pertenecería cada muestra.

El segundo enfoque, consiste en la predicción del valor exacto de cada componente. Este fue el enfoque que finalmente se utilizó, principalmente debido a que las cantidades consideradas para establecer franjas podían ser relativas a los valores de contaminación propios de la ciudad o en forma absoluta para lo que se recomienda según organismos y estudios enfocados en calidad habitacional de una ciudad.

El algoritmo que se decidió aplicar es el Support Machine Regressor (SVR^3), debido a la cantidad de muestras y variables.

Este método construye un hiper plano que, en base a una penalización o costo, determina un radio o margen. Este determinará las muestras sobre las que se "apoyará" el tubo (support vectors)

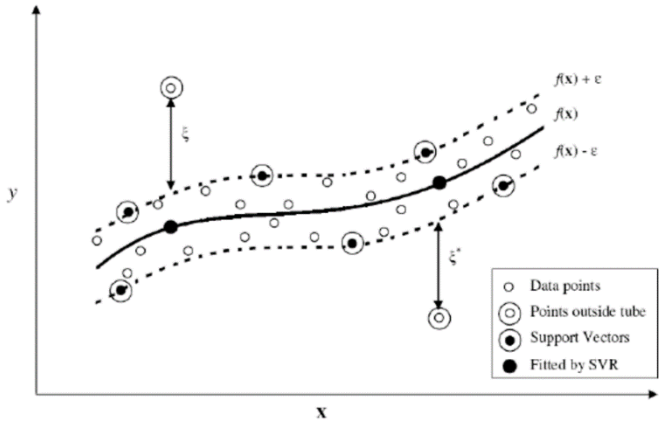

Figura 6

Para identificar los mejores hiper parámetros para el modelo se utilizará el método **GridSearch** que permitirá probar distintos hiper parámetros y a la vez nos permitirá realizar la cross validation, esto es dividir el modelo en distintos folds que harán de set de entrenamiento y prueba en forma sistemática permitiendo validar el entrenamiento en con un fold de prueba en cada iteración.

La muestra de testeo que se utilizará será del 15% del dataset.

No obstante, a fines de no limitar o sesgar el análisis se procedió a evaluar el resultado de otros algoritmos. Los resultados se pueden comparar en la **Tabla 1**.

Modelo	CO		NO_2		PM_{10}	
	MAE	R^2	MAE	R^2	MAE	R^2
Linear Reg.	0.10	0.277	5.16	0.218	7.54	0.058
KNN	0.13	-0.188	6.15	-0.102	-	-
SVR	0.098	0.297	4.89	0.24	6.85	0.138

Tabla 1

Basándose en estos resultados se podría decir que si bien las variables consideradas no son las mejores para predecir estos valores el error no es tan grande dentro de las dimensiones de cada variable.

Además, entrando en detalle de las predicciones se puede observar que la mayor acumulación de los errores de las predicciones (naranja) radica en los "picos" o valores extremos que el modelo no logra alcanzar (**Figuras 7, 8 y 9**).

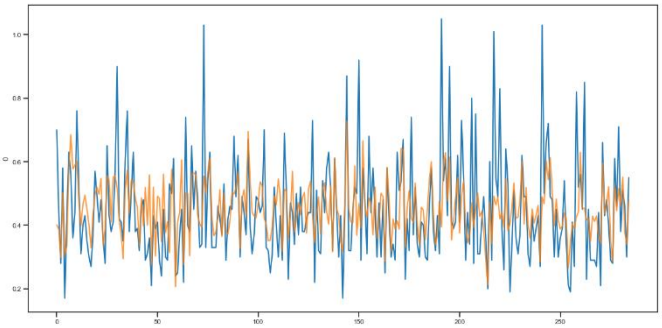

Figura 7

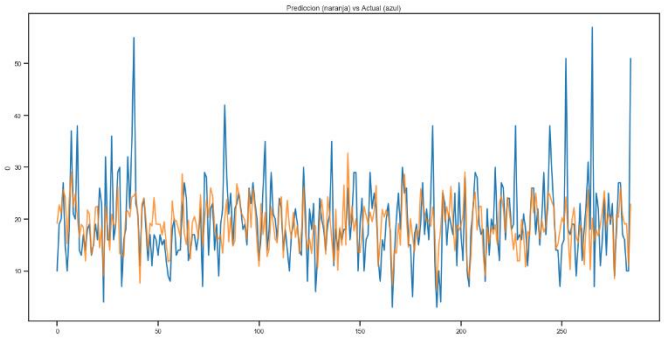

Figura 8

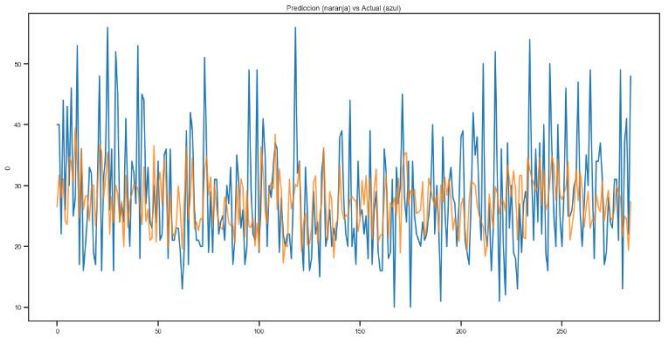

Figura 9

Los altos niveles de variación entre las muestras hacen comprensible este comportamiento.

Por último, se realizó la predicción de los valores faltantes que habían sido removidos durante la limpieza del dataset. Para esto se volvió a entrenar el modelo con los mejores parámetros. En las figuras 10, 11 y 12 pueden observarse de forma grafica los resultados.

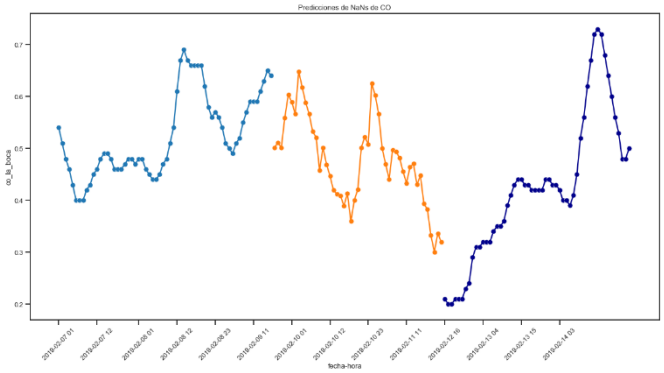

Figura 10

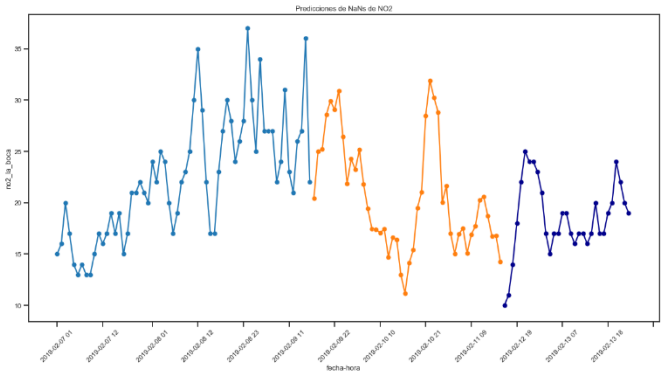

Figura 11

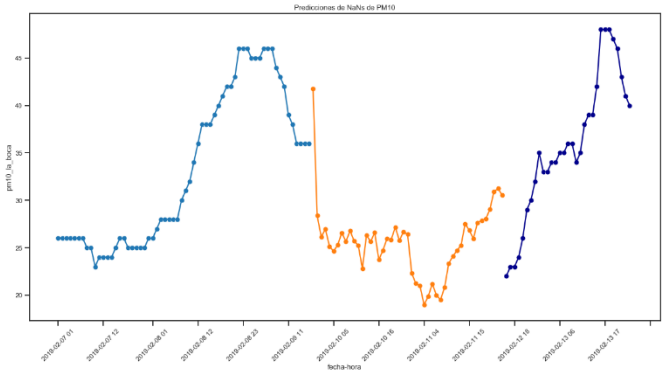

Figura 12

Puede observarse que la gráfica del modelo de NO_2 y CO es da resultados mucho más acordes a lo que cabria esperar a la variación entre los intervalos temporales de lo que demuestra la medición de PM_{10}

4 CONCLUSIONES Y DISCUSION

A partir de los resultados obtenidos logramos definir que a partir de las consideradas se puede estimar con una mínima aproximación de la presencia de CO y NO_2, y en menor medida la de PM_{10}.

Tras haber indagado en la gran diferencia en cuanto a la predictibilidad de las PM_{10} se logró identificar que en las proximidades de la estación de medición se estuvieron llevando a cabo obras del *"Paseo del bajo"* durante todo el periodo de análisis de que consideró este informe y estas obras de construcción pudieron haber alterado los valores normales de material particulado en las mediciones.

Dicho esto, los autores del presente informe sostienen que considerando el propósito inicial del mismo los valores obtenidos permitirían una primera aproximación que, con muestras que abarquen un intervalo de tiempo más elevado, y/o mayor diversidad respecto a los puntos de medición de las mismas se podría aumentar la precisión y lograr realizar el mapa de calidad de aire explicado previamente que sería sumamente útil para las gestiones urbanísticas de la ciudad de buenos aires y el cuidado de la salud de sus ciudadanos.

5 REFERENCIAS

1. Nota del diario de Madrid sobre ganadores de la hackatón llevada a cabo por la empresa SAS https://diario.madrid.es/blog/notas-de-prensa/un-modelo-predictivo-medioambiental-gana-el-hackaton-convocado-por-ayuntamiento-y-sas-espana/
2. Información disponible en el portal de datos del Gobierno de la Ciudad de Buenos Aires
3. A tutorial on Principal Components Analysis, Lindsay I Smith, February 2002
4. Support Vector Regression Machines, Harris Drucker. Chris J.C. Burges, Linda Kaufman, Alex Smola Vladimir Vapoik

6 ADJUNTOS

Script mediante el cual se mergearon las tablas https://github.com/EVGlovo/ClusterAI/blob/master/Mergeo%20de%20tablas%20-%20Trabajo%20Calidad%20de%20Aire.ipynb

Repositorio de GitHub donde se puede encontrar el EDA realizado

https://github.com/EVGlovo/ClusterAI/blob/master/clusterai_ezequiel_vannucchi_eda.ipynb

Repositorio de GitHub onde se puede encontrar el script con el modelo utilizado

https://github.com/EVGlovo/ClusterAI/blob/master/clusterai_ezequiel_vannucchi_machine_learning.ipynb

7 RECONOCIMIENTOS

Agradecemos a los profesores de ciencia de datos Martin Palazzo, Nicolas Aguirre y Agustin Velazques por brindarnos el espacio y la oportunidad de realizar este trabajo y facilitarnos la posibilidad de adquirir habilidades esenciales para el desarrollo profesional en la actualidad, así como la piedra fundamental para los desarrollos venideros.