Departamento de Matemáticas y Física

Desarrollo a productivo de un modelo predictivo de calidad del aire

---

MACHINE LEARNING OPERATIONS BOOTCAMP

Presenta: **DANIEL MARTINEZ ESCOBOSA**

Guadalajara, Jalisco. marzo de 2024.

TABLA DE CONTENIDO

LISTA DE FIGURAS

LISTA DE TABLAS

1. INTRODUCCIÓN

1.1. Problema

Abordar un desafío identificando las razones fundamentales que lo justifican. Se resolverá un problema común de aprendizaje automático (problema de regresión) utilizando un data set proporcionado y aplicando lo aprendido en el entrenamiento.

1.2. Objetivos

1.2.1. Objetivo General

Los objetivos generales del programa radican en:

- Aplicar conocimientos teóricos en un contexto práctico.
- Demostrar competencia en los temas cubiertos del programa.
- Demostrar habilidades para resolver problemas en escenarios del mundo real utilizando un data set.

1.2.2. Objetivos Específicos

1. Analizar y comprender el data set proporcionado mediante un Análisis Exploratorio de Datos (EDA por sus siglas en inglés).
2. Determinar la pregunta que deseas hacer con un modelo de aprendizaje automático.
3. Identificar por qué se necesita una estrategia de MLOps para este data set (Módulo 1).
4. Diseñar la arquitectura del pipeline para esta nueva iniciativa de aprendizaje automático.
5. Crear un modelo de referencia para resolver tareas asociadas de predicción (clasificación, regresión, etc.) y que responda a la pregunta que te hiciste. Este modelo no necesita tener una precisión, recall o F1-Score altos; sino de crear un modelo rápido para iterar (Módulo 4). Elegir sus características y los hiperparámetros iniciales para este modelo.
6. Configurar la estructura correcta del modelo con la idea de dejarlo listo para implementarse (Módulo 6).
7. Crear nuevas versiones del modelo, que incluyan cambios en las características o ajustes de hiperparámetros para la reproducibilidad y el seguimiento experimental (Módulo 7).
8. Implementar un modelo en su entorno local usando contenedores y planificar el reentrenamiento, drift, redeploy, escalado y monitoreo (Módulo 8).
9. Demostrar estrategias de prueba y versionamiento (Módulos 7 y 8).

2. MARCO TEÓRICO/CONCEPTUAL

2.1. Técnicas estadísticas y modelos predictivos de regresión

2.1.1. Análisis de correlación

La correlación, también conocida como coeficiente de correlación lineal (de Pearson), es una medida de regresión que pretende cuantificar el grado de variación conjunta entre dos variables.

Para estudiar la relación lineal existente entre dos variables continuas es necesario disponer de parámetros que permitan cuantificar dicha relación. Uno de estos parámetros es la covarianza, que indica el grado de variación conjunta de dos variables aleatorias:

$$Cov(X,Y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{N-1} \quad (1)$$

siendo $\bar{x}$ e $\bar{y}$ las medias de cada variable y x_i e y_i el valor de las variables para la observación i.

La covarianza depende de las escalas en que se miden las variables estudiadas, por lo tanto, no es comparable entre distintos pares de variables. Para poder hacer comparaciones se estandariza la covarianza, generando lo que se conoce como coeficientes de correlación. Existen diferentes tipos, de entre los que destacan el coeficiente de Pearson, Rho de Spearman y Tau de Kendall. Todos ellos varían entre +1 y -1. Siendo +1 una correlación positiva perfecta y -1 una correlación negativa perfecta.

La correlación de Pearson, es la más común. Funciona bien con variables cuantitativas que tienen una distribución normal (e incluso se le considera robusto a pesar de la falta de normalidad), aunque es más sensible a los valores extremos que las otras dos alternativas.

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} \quad (2)$$

2.1.2. Análisis de Componentes Principales (PCA)

El Análisis de Componentes Principales (PCA por sus siglas en inglés) es un método estadístico de aprendizaje no supervisado que permite simplificar la complejidad de espacios muestrales con muchas dimensiones a la vez que conserva su información.

Las componentes principales se obtienen por combinación lineal de las variables originales. Se pueden entender como nuevas variables obtenidas al combinar de una determinada forma las variables originales. La primera componente principal de un grupo de variables $(X_1, X_2, \dots, X_p)$ es la combinación lineal normalizada de dichas variables que tiene mayor varianza:

$$Z_1 = \varphi_{11}X_1 + \varphi_{21}X_2 + \cdots + \varphi_{p1}X_p \quad (3)$$

Y que entendido de forma gráfica se puede observar la Figura 9.

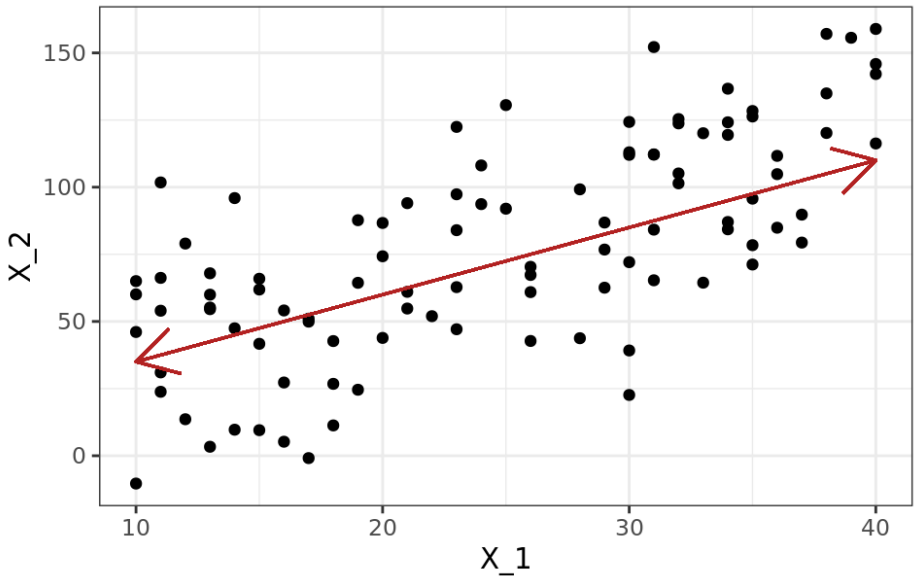


Figura 1. Componente principal Z1 dirección mayor varianza

Que la combinación lineal sea normalizada implica que:

$$\sum_{j=1}^{p} \varphi_{j1}^2 = 1 \quad (4)$$

Los términos φ_{11}, …, φ_{1p} reciben en el nombre de loadings y son los que definen a la componente. Por ejemplo, φ_{11} es el loading de la variable X1 de la primera componente principal. Los loadings pueden interpretarse como el peso o importancia que tiene cada variable en cada componente y, por lo tanto, ayudan a conocer la información que recoge cada una de las componentes.

Para conocer cuanta información es capaz de capturar cada una de las componentes principales obtenidas, se recurre a la proporción de varianza explicada por cada una de ellas. Asumiendo que las variables se han normalizado para tener media cero, la varianza total presente en el set de datos se define como:

$$\sum_{j=1}^{p} Var(X_j) = \sum_{j=1}^{p} \frac{1}{n} \sum_{i=1}^{n} x_{ij}^2 \quad (5)$$

y la varianza explicada por la componente m es:

$$\frac{1}{n} \sum_{i=1}^{n} z_{im}^2 = \frac{1}{n} \sum_{i=1}^{n} \left(\sum_{j=1}^{p} \varphi_{jm} x_{ij} \right)^2 \quad (6)$$

Tanto la proporción de varianza explicada, como la proporción de varianza explicada acumulada son dos valores de gran utilidad a la hora de decidir el número de componentes principales a utilizar para reducción de dimensionalidad.

2.1.3. Regresión lineal múltiple

La regresión lineal múltiple permite generar un modelo lineal en el que el valor de la variable dependiente o respuesta (Y) se determina a partir de un conjunto de variables independientes llamadas predictores (X1, X2, X3…). Los modelos de regresión múltiple pueden emplearse para predecir el valor de la variable dependiente o para evaluar la influencia que tienen los predictores sobre ella.

Los modelos lineales múltiples siguen la siguiente ecuación:

$$Y_i = (\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_n X_{ni}) + e_i \quad (7)$$

En donde:
β_0: es la ordenada en el origen, el valor de la variable dependiente Y cuando todos los predictores son cero.
β_i: se conocen como coeficientes parciales de regresión. Representan el efecto promedio que tiene el incremento en una unidad de la variable predictora X_i sobre la variable dependiente Y, manteniéndose constantes el resto de las variables.
e_i: es el residuo o error, la diferencia entre el valor observado y el estimado por el modelo.

2.1.4. Árboles de decisión

Un árbol de decisión es un conjunto de oraciones incrustadas de la forma "si, entonces" que dividen el conjunto de datos de los predictores. Las particiones en el conjunto de predictores se utilizan para estimar la salida del modelo como se ejemplifica en la Figura 10.

La estructura de un árbol se basa principalmente en 3 puntos principales:

- El predictor o predictores que se utilizarán y el punto de partición del conjunto de datos.
- La profundidad o complejidad del árbol.
- La ecuación de predicción en los nodos terminales u hojas del árbol.

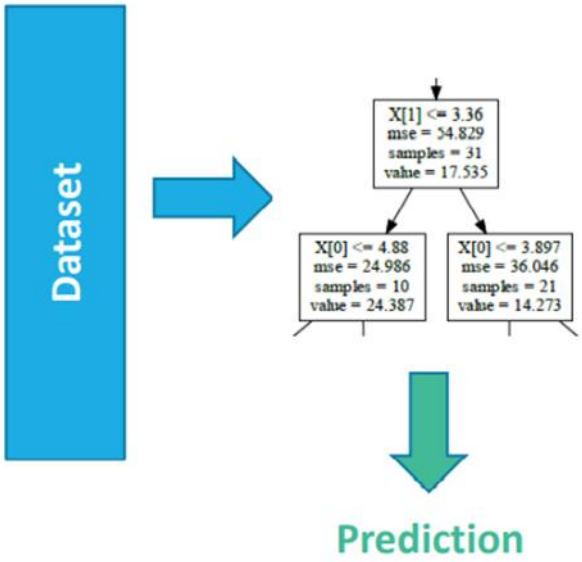


Figura 2. Esquema del árbol de decisión

Para un problema de regresión, se requiere encontrar o proponer un modelo para lograr un mapeo entre dos conjuntos de datos.

$$Y = f(X) \quad (8)$$

donde Y y X son el conjunto de variables de salida o a estimar, y el conjunto de predictores o variables de entrada, respectivamente. La función $f(\cdot)$ será aproximada por el árbol de decisión.

Para la construcción de un árbol de decisión, se considera el conjunto de datos completo S, donde $X, Y \in S$. El conjunto de datos S se divide mediante una búsqueda exhaustiva de todas las particiones posibles para crear dos subconjuntos de datos $S1$ y $S2$, de tal forma que se minimiza un criterio de varianza conjunta, $L(x_{split})$.

$$L(x_{pj,split}) = \frac{1}{n_1}\sum_{i \in S_1}(y_i - \bar{y}_1)^2 + \frac{1}{n_2}\sum_{i \in S_2}(y_i - \bar{y}_2)^2 \quad (9)$$

Donde, $x_{pj,split}$ es un valor de partición en el predictor pj y los parámetros $\bar{y}_1$ y $\bar{y}_2$ son los promedios de las muestras de la variable de salida pertenecientes a los respectivos subconjuntos $S1$ y $S2$.

El valor de la función $L(x_{px,split})$, depende de la mejor $x_{px,split}$, que se selecciona para minimizar la varianza en la variable de salida Y. La búsqueda de la mejor partición debe realizarse de forma exhaustiva en todos los predictores x_{pj}, y solo se elige una partición de entre todos los predictores.

Este procedimiento se realiza recursivamente en los subconjuntos $S1$ y $S2$.

La función de predicción en cada hoja del árbol se define como el valor promedio de la salida muestras del subconjunto de datos Sj correspondientes a la j-ésima rama o nodo final. La predicción $\hat{y}_i$ correspondiente a una muestra x_i se define como:

$$\hat{y}_{i.j} = \bar{y}_j \quad (10)$$

2.1.4.1. Bolsa de árboles de decisión

El ensamblaje de modelos es una estrategia que intenta resolver el problema de la solución subóptima de árboles de decisión.

El método "Bagging" es relativamente simple ya que utiliza "bootstrapping" junto con diferentes técnicas de regresión para construir un modelo ensamblado. Cada modelo en el empaquetado genera una predicción para la misma muestra, y todas las predicciones son promediadas para dar una predicción definitiva (ver Figura 11). De esta forma, al combinar sus resultados, unos errores se compensan con otros, reduciéndose la varianza en la predicción y haciendo la predicción más estable y con poder de generalización.

La construcción de una bolsa de árboles se hace de forma sencilla:

1. Generar un subconjunto de datos a partir de los datos originales.
2. Entrenar un árbol de decisiones no podado basado en el subconjunto de datos.
3. Repetir los pasos 1 y 2 hasta que se formen m árboles de decisión.

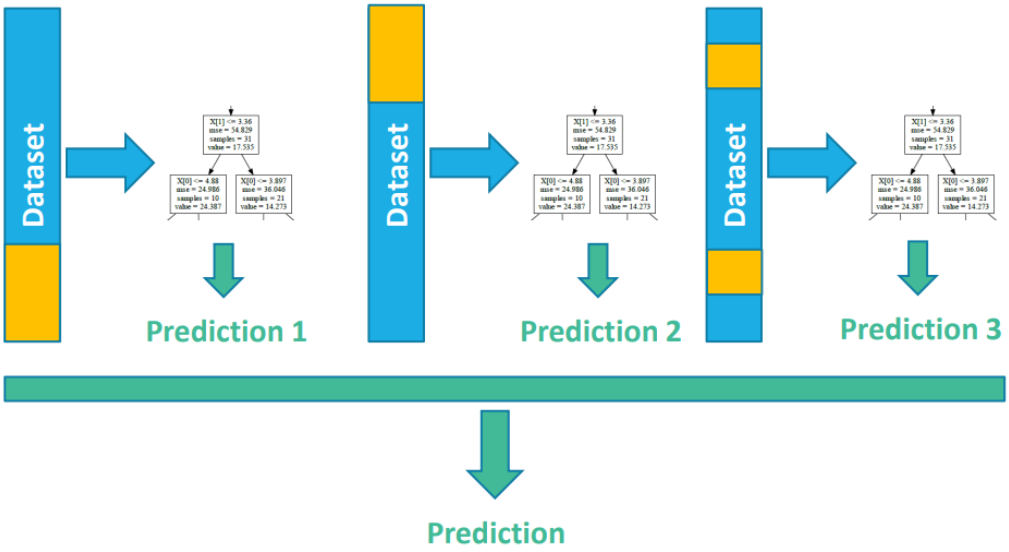


Figura 3. Esquema del modelo de bolsa de árboles de decisión

2.1.5. Red Neuronal

Una red neuronal artificial es un modelo altamente no lineal y muy versátil que se usa comúnmente para resolver problemas de regresión o clasificación.

La forma en la que una neurona artificial procesa una señal se representa en dos etapas, tal como se muestra en la Figura 12 y se explica a continuación:

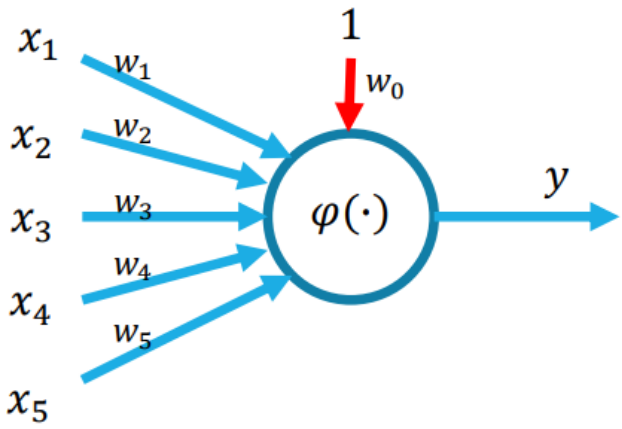


Figura 4. Representación neurona artificial

1. Combinación de las entradas. Todas las señales que recibe la neurona deben ser consideradas para generar una respuesta. Cada señal recibida tiene asociada una

constante w, que permite manipular la importancia de cada señal recibida o el aporte de información de cada entrada. Su representación matemática se da por la siguiente ecuación:

$$v = w_0 + w_1x_1 + w_2x_2 + w_3x_3 + w_4x_4 + w_5x_5 \quad (11)$$

2. Respuesta neuronal. Considerando todas las señales de entrada, la neurona genera una señal de respuesta $y = \varphi(v)$ por medio de una función de activación, la cual es diferenciable y ayuda a resolver el problema de optimización en el entrenamiento. Las funciones de activación más comunes son:

$$\text{Lineal: } \varphi(v) = v \quad (12)$$

$$\text{Sigmoidal: } \varphi(v) = \frac{1}{1+e^{-v}} \quad (13)$$

$$\text{Tangente hiperbólica: } \varphi(v) = \tanh(v) \quad (14)$$

Ahora bien, partiendo de la teoría de una neurona artificial, una red neuronal es la interconexión de múltiples neuronas para resolver un problema complejo que una sola neurona no podría. Su organización, como se visualiza en la Figura 13, se realiza en capas para simplificar el análisis: de entrada, ocultas y de salida.

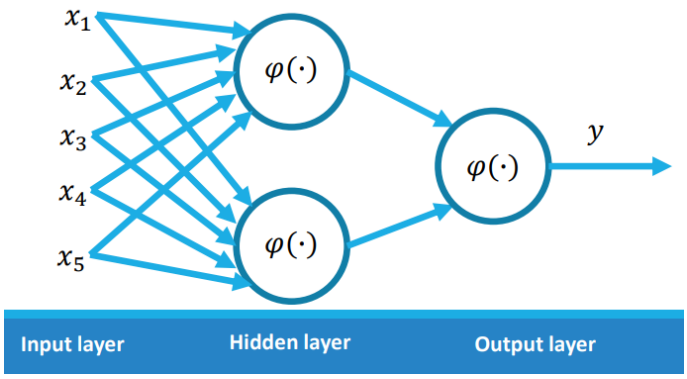


Figura 5. Capas de una red neuronal

Intuitivamente, se puede ver que una red neuronal proporciona un modelo no lineal y esto se puede ajustado por medio de los pesos de cada capa.

Para un problema de regresión, se usa que la función de activación en la capa de salida sea la lineal, así mismo la evaluación de una red neuronal se realiza mediante una función de error o función de costo J.

$$J = \frac{1}{2n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad (15)$$

Donde, y_i es la salida deseada, $\hat{y}_i$ es la salida de la red neuronal para una entrada x_i y n es el número de salidas o neuronas en la capa de salida.

Por tanto, la evaluación de la red neuronal para un conjunto de muestras m se puede escribir como:

$$J = \frac{1}{2nm}\sum_{j=1}^{m}\sum_{i=1}^{n}\left(y_{i,j} - \hat{y}_{i,j}\right)^2 \quad (16)$$

Se considera que un modelo neuronal ya tiene aprendido el comportamiento de los datos a ajustar cuando se tiene un valor mínimo de la función J. Para lograr esto es necesario encontrar los pesos W que minimicen la evaluación de J para todas las muestras de un conjunto de datos.

Ahora bien, cuando se aplica la función de costo J con un modelo neuronal, los pesos W tienen una relación no lineal a la salida y esto hace que J sea una función no convexa. Es decir, no hay una única solución W que minimiza la función de costo J. Como ayuda ante ello se utiliza un algoritmo de aprendizaje cuya función es la minimización de J modificando los parámetros W de la red neuronal.

Durante el proceso de aprendizaje o entrenamiento de una red neuronal, la actualización de los pesos es iterativamente siguiendo la siguiente regla:

$$W_k = W_{k-1} + \Delta W_{k-1} \quad (17)$$

Los algoritmos de entrenamiento de redes neuronales más conocidos son:

a) Gradiente descendente
b) Método de Newton
c) Gauss-Newton
d) Levenberg-Marquart

3. DESARROLLO METODOLÓGICO

3.1. *Procedencia y método de obtención de datos*

El conjunto de datos contiene 9358 casos de respuestas promediadas por hora de una serie de cinco sensores químicos de óxido metálico integrados en un dispositivo multisensor químico de calidad del aire. El dispositivo estaba ubicado en el campo en una zona significativamente contaminada, al nivel de la carretera, dentro de una ciudad italiana. Los datos se registraron desde marzo de 2004 hasta febrero de 2005 (un año), lo que representa las grabaciones más largas disponibles de forma gratuita de las respuestas de dispositivos de sensores químicos de calidad del aire desplegados en el campo. Ground Truth promedió las concentraciones horarias de CO, hidrocarburos no metánicos, benceno, óxidos de nitrógeno total (NOx) y dióxido de nitrógeno (NO2) y fueron proporcionadas por un analizador certificado de referencia ubicado en el mismo lugar. Están presentes evidencias de sensibilidades cruzadas, así como desviaciones tanto de conceptos como de sensores, como se describe en De Vito et al., Sens. And Act. B, vol. 129,2,2008 (cita requerida) que eventualmente afecta las capacidades de estimación de concentración de los sensores. Los valores faltantes están etiquetados con el valor -200. Este conjunto de datos se puede utilizar exclusivamente con fines de investigación. Los fines comerciales están totalmente excluidos.

3.2. *Análisis Exploratorio de Datos (EDA)*

Haciendo un análisis exploratorio de la información presente en el dataset, se encontraron los siguientes hallazgos relevantes:

- En términos de análisis de contaminantes en el tiempo, la variable objetivo llamada CO(GT)" presentó los niveles más altos durante el periodo de tiempo entre noviembre 2004 a enero 2005, en donde coincide con un decrecimiento en la temperatura como se aprecian en las dos figuras siguientes:

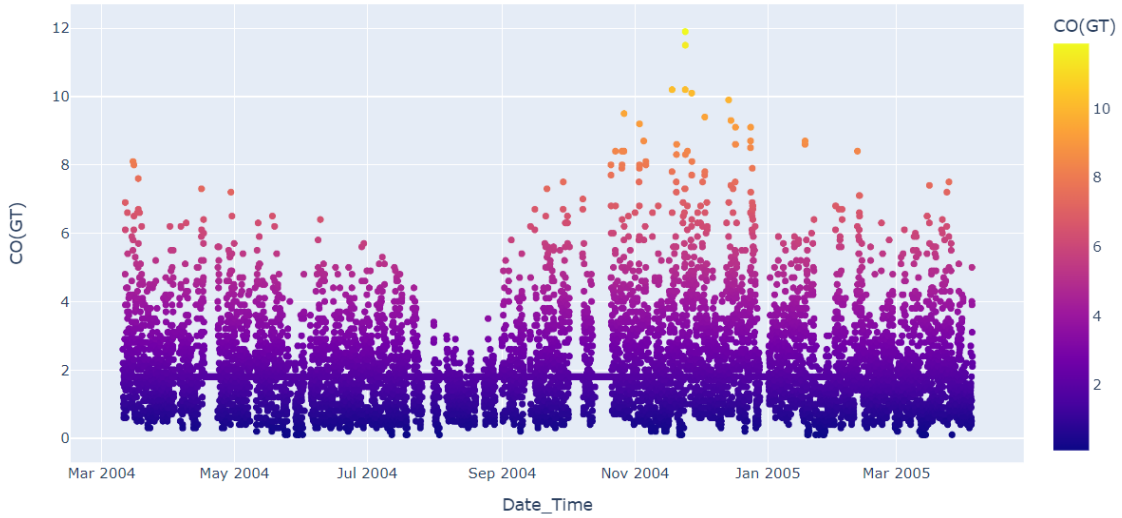


Figura 6. Niveles de Monóxido de Carbono CO(GT) en el tiempo

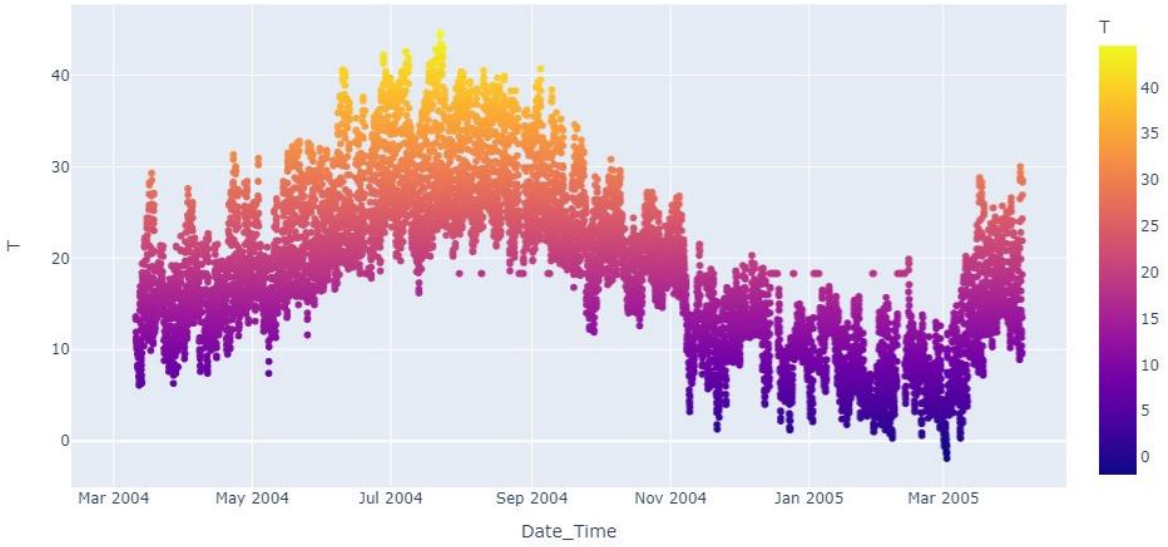


Figura 7. Temperatura “T” en grados Celsius durante el año

- Al estudiar los histogramas previo a tratamiento de datos, se detectó un sesgo moderado en las variables “NMHC(GT)” y “NOx(GT)”, cuyos valores de sesgo “Skew” fueron >1.5 pero <2, o <-1.5 pero >-2 (criterio empírico); mientras que las variables con la distribución más cercana a una normal, fueron “AH”, “RH”, “T” y “PT08.S4(NO2)” cayendo dentro del rango -0.5<Skew<0.5.

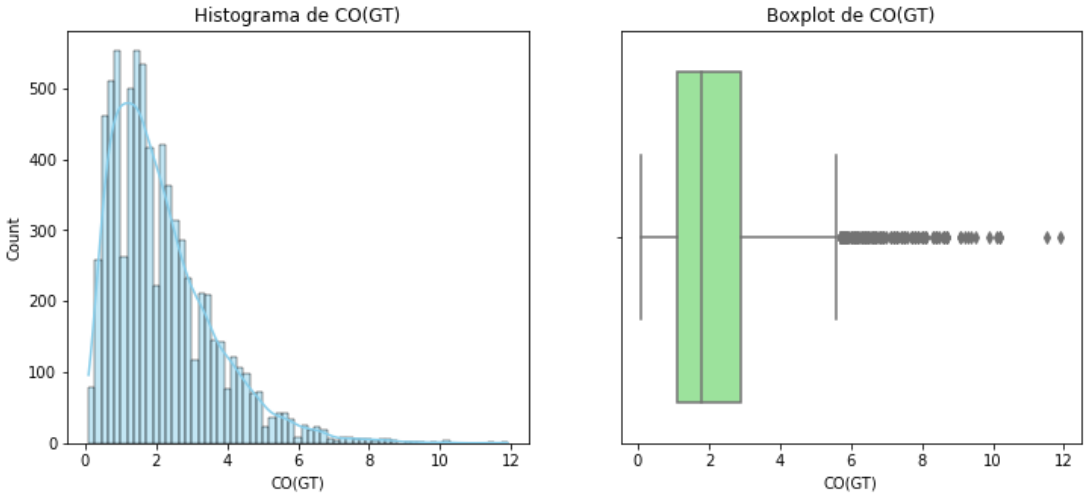


Figura 8. Estudio de distribuciones y presencia de outliers

- Con la ayuda de un análisis de correlación mediante mapas de Seaborn, se detectó que la variable “PT08.S3(NOx)” presentó alto índice de correlación entre sus pares de variables, tal como se aprecia en la siguiente figura:

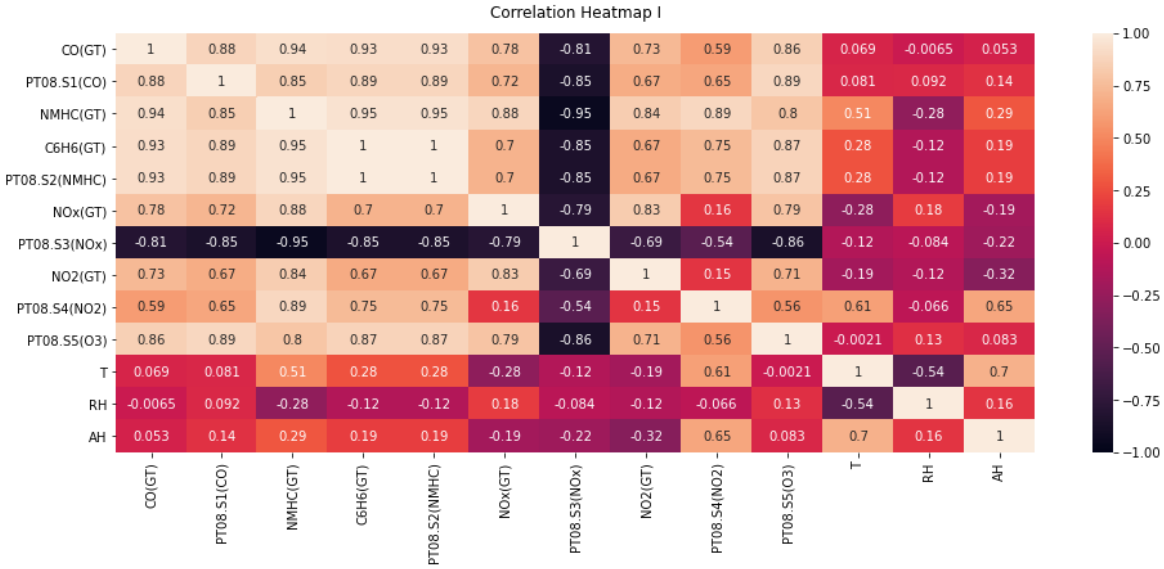


Figura 9. Mapa de calor de correlación entre pares de variables

- El gráfico de la figura representa cómo se distribuye el "total" de las medias (considerado como la suma de todas ellas) entre las distintas variables, destacando la variable de “PT08.S4(NO2)”.

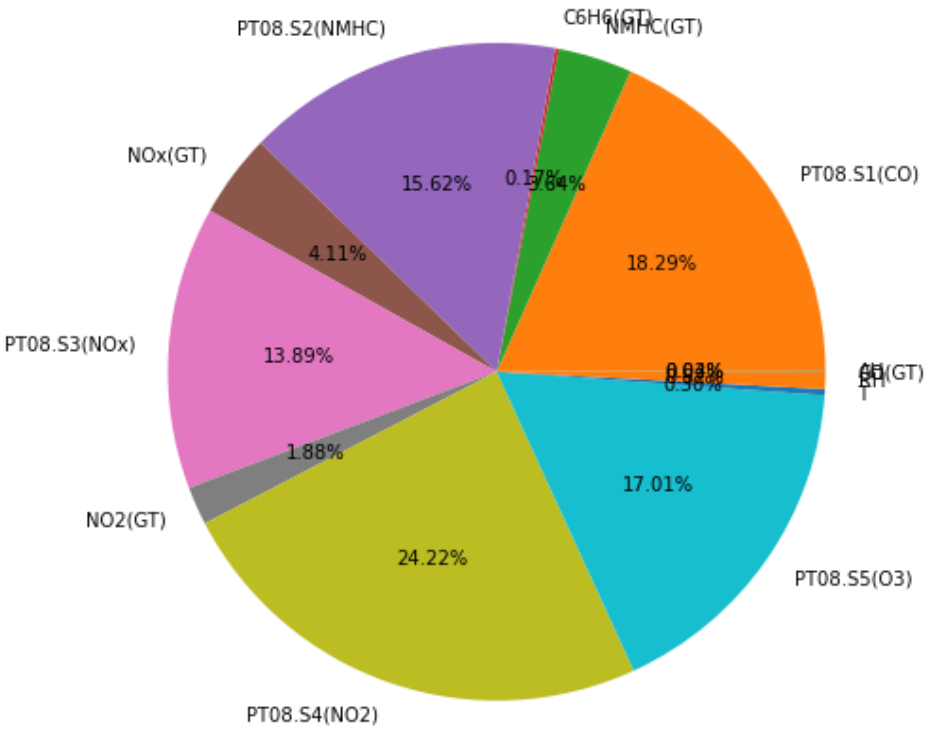


Figura 10. Distribución de medias

3.3. *Determinar Pregunta del Modelo de Aprendizaje*

Pensando hipotéticamente que se trata de un gas difícil de medir por el alto costo del sensor que capta su valor, predecir el valor numérico del gas CO(GT) que se tendría basado en los valores de los otros componentes (features).

3.4. Justificación estrategia de MLOps para data set

El dataset de calidad del aire que se utiliza es una fuente valiosa de datos para comprender y predecir la calidad del aire en función de varias variables ambientales. A continuación, se explican algunos puntos por los que este dataset requiere una estrategia de MLOps:

- **Tamaño del dataset:** El dataset contiene registros de sólo un año, pero pensando en múltiples años o incluso en hipotéticas mediciones futuras en menor periodicidad -por ejemplo, mediciones por minuto-, podría manejarse gran volumen de datos. Esto implica desarrollar estrategias de gestión de datos eficiencias como técnicas de almacenamiento distribuido, particionamiento de

datos y gestión de versiones para asegurar la reproducibilidad de los experimentos.

- **Preprocesamiento de datos:** El dataset de estudio contiene datos incompletos, ruidosos e inconsistentes que necesitan ser preprocesados antes de ser utilizados en modelos de aprendizaje automático. Mediante MLOps se incluiría la automatización de pipelines de preprocesamiento de datos para garantizar la consistencia y la escalabilidad del proceso.

- **Selección de características:** Es necesario identificar las características más relevantes para predecir la calidad del aire es crucial. Con MLOps se puede incluir la automatización de la selección de características utilizando técnicas como la ingeniería de características automática o la selección de modelos.

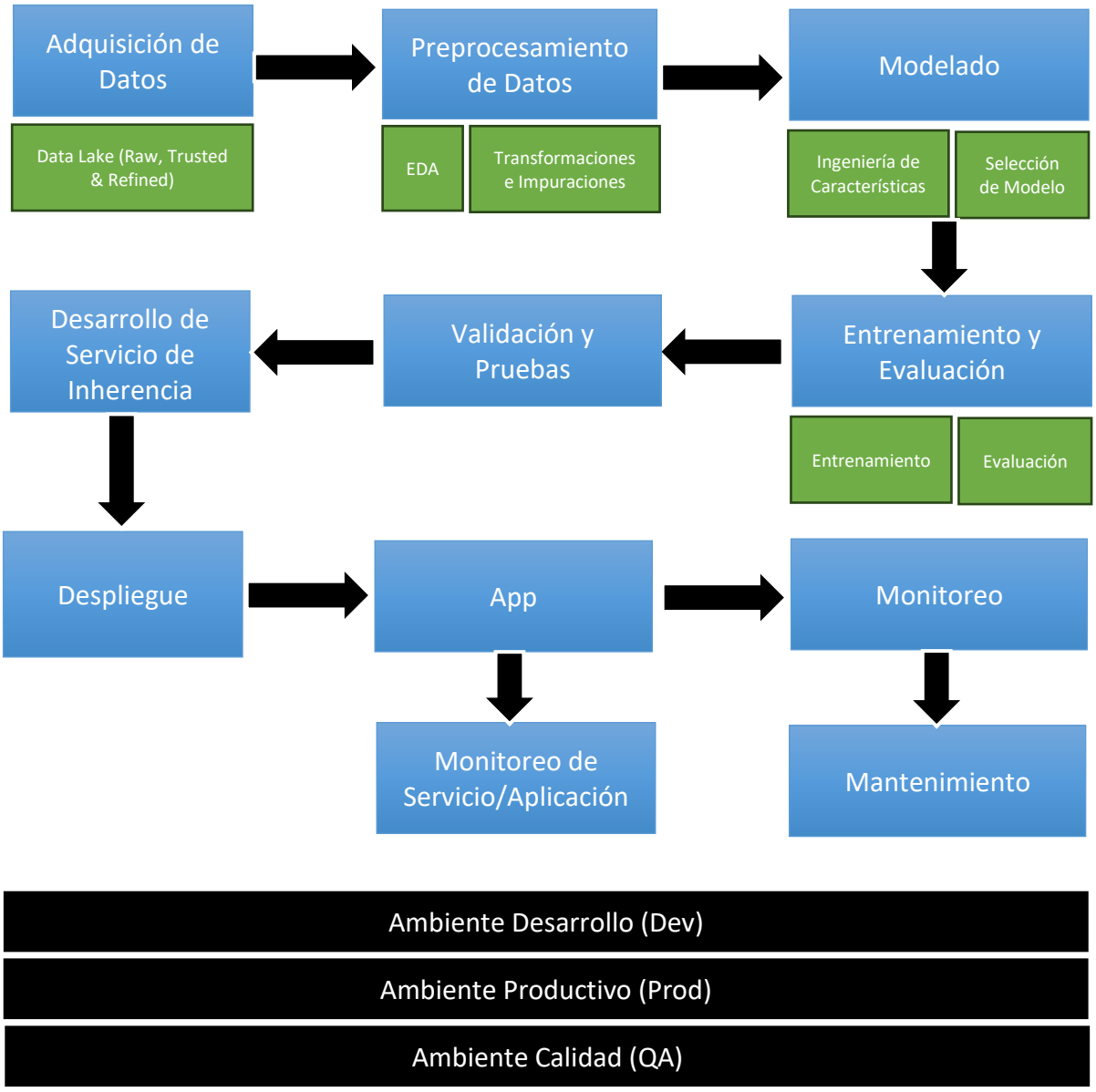
- **Entrenamiento de modelos:** Implementar pipelines de entrenamiento de modelos que sean escalables y reproducibles es esencial. Esto implica la configuración de infraestructura de cómputo escalable, la implementación de pipelines de entrenamiento utilizando herramientas como TensorFlow Extended (TFX) o Apache Airflow, y la gestión de recursos para garantizar la eficiencia y la calidad de los modelos.

- **Evaluación y monitorización de modelos en producción:** Una vez que el modelo esté en producción, es importante monitorear su rendimiento y calidad de predicción. Con MLOps se podría incluir en la implementación sistemas de monitorización en tiempo real, alertas automáticas para detectar anomalías en los datos o en el rendimiento del modelo, y la retroalimentación continua para mejorar el modelo con nuevos datos.

- **Despliegue y gestión de modelos en producción:** Implementar estrategias de despliegue de modelos robustas y seguras es fundamental. Esto implica la automatización del proceso de despliegue utilizando contenedores como Docker, la gestión de versiones de modelos, la configuración de infraestructura de producción escalable y la implementación de prácticas de seguridad para proteger los modelos y los datos.

En resumen, la aplicación de estrategias de MLOps en este proyecto ayuda a gestionar eficientemente todas las etapas del ciclo de vida del modelo de aprendizaje automático, desde la gestión de datos hasta el despliegue y monitorización en producción, garantizando así la escalabilidad, reproducibilidad y calidad del modelo.

3.5. *Arquitectura Pipeline Iniciativa de Aprendizaje Automático*



3.6. *Modelado de Referencia*

Se pretenden desarrollar múltiples modelos, sin embargo, para esta entrega se hace el desarrollo de un modelo de regresión multivariado utilizando un split de 80% de entrenamiento y 20% de prueba de los datos preprocesados en pasos anteriores y para el cual se hizo uso de la técnica de validación cruzada para garantizar que los resultados obtenidos fueran independientes de la partición entre datos de entrenamiento y prueba. Mientras que los KPIs propuestos son el coeficiente de determinación (R^2) y el RMSE.

A continuación, los resultados arrojados del ejercicio:

Modelo	**LR**	**Modelo 2**	**…**	**Modelo N**
Train	0.7260	Pendiente	Pendiente	Pendiente
Test	0.6758	Pendiente	Pendiente	Pendiente

Tabla 1. KPI Coeficiente de Determinación

Modelo	**LR**	**Modelo 2**	**…**	**Modelo N**
Train	0.5577	Pendiente	Pendiente	Pendiente
Test	0.6314	Pendiente	Pendiente	Pendiente

Tabla 2. Valores de RMSE

4. RESULTADOS Y DISCUSIÓN

4.1. Resultados Modelos

Por desarrollar una vez se realicen más modelos predictivos.

5. CONCLUSIONES

5.1. *Conclusiones*

Por desarrollar en siguientes entregables.

BIBLIOGRAFÍA

[1] *J. Amat, "Correlación lineal y regresión lineal" [Online document] Jun. 2016, Available at HTTP:* https://www.cienciadedatos.net/documentos/24_correlacion_y_regresion_lineal

[2] *J. Amat, "Análisis de componentes principales (Principal Component Analysis, PCA) y t-SNE" [Online document] Jun. 2017, Available at HTTP:* https://www.cienciadedatos.net/documentos/35_principal_component_analysis

[3] *J. Amat, "Introducción a la regresión lineal múltiple" [Online document] Jul. 2016, Available at HTTP:* https://www.cienciadedatos.net/documentos/25_regresion_lineal_multiple.html

[4] M. Kuhn y K. Johnson, "Decision Trees and Neural Networks" en Applied Predictive Modeling, NY: Springer New York, 2013, chapter 14.

APÉNDICE A. GLOSARIO DE VARIABLES

A continuación, se presentan las variables contenidas en el dataset estudiado y que se abordó para este trabajo:

Variable Name	Role	Type	Description	Units	Missing Values
Date	Feature	Date			no
Time	Feature	Categorical			no
CO(GT)	Feature	Integer	True hourly averaged concentration CO in mg/m^3 (reference analyzer)	mg/m^3	no
PT08.S1(CO)	Feature	Categorical	hourly averaged sensor response (nominally CO targeted)		no
NMHC(GT)	Feature	Integer	True hourly averaged overall Non Metanic HydroCarbons concentration in microg/m^3 (reference analyzer)	microg/m^3	no
C6H6(GT)	Feature	Continuous	True hourly averaged Benzene concentration in microg/m^3 (reference analyzer)	microg/m^3	no
PT08.S2(NMHC)	Feature	Categorical	hourly averaged sensor response (nominally NMHC targeted)		no
NOx(GT)	Feature	Integer	True hourly averaged NOx concentration in ppb (reference analyzer)	ppb	no
PT08.S3(NOx)	Feature	Categorical	hourly averaged sensor response (nominally NOx targeted)		no
NO2(GT)	Feature	Integer	True hourly averaged NO2 concentration in microg/m^3 (reference analyzer)	microg/m^3	no
PT08.S4(NO2)	Feature	Categorical	hourly averaged sensor response (nominally NO2 targeted)		no
PT08.S5(O3)	Feature	Categorical	hourly averaged sensor response (nominally O3 targeted)		no
T	Feature	Continuous	Temperature	°C	no
RH	Feature	Continuous	Relative Humidity	%	no

AH	Feature	Continuous	Absolute Humidity		no