UEMC
Universidad Europea Miguel de Cervantes

Trabajo Fin de Máster

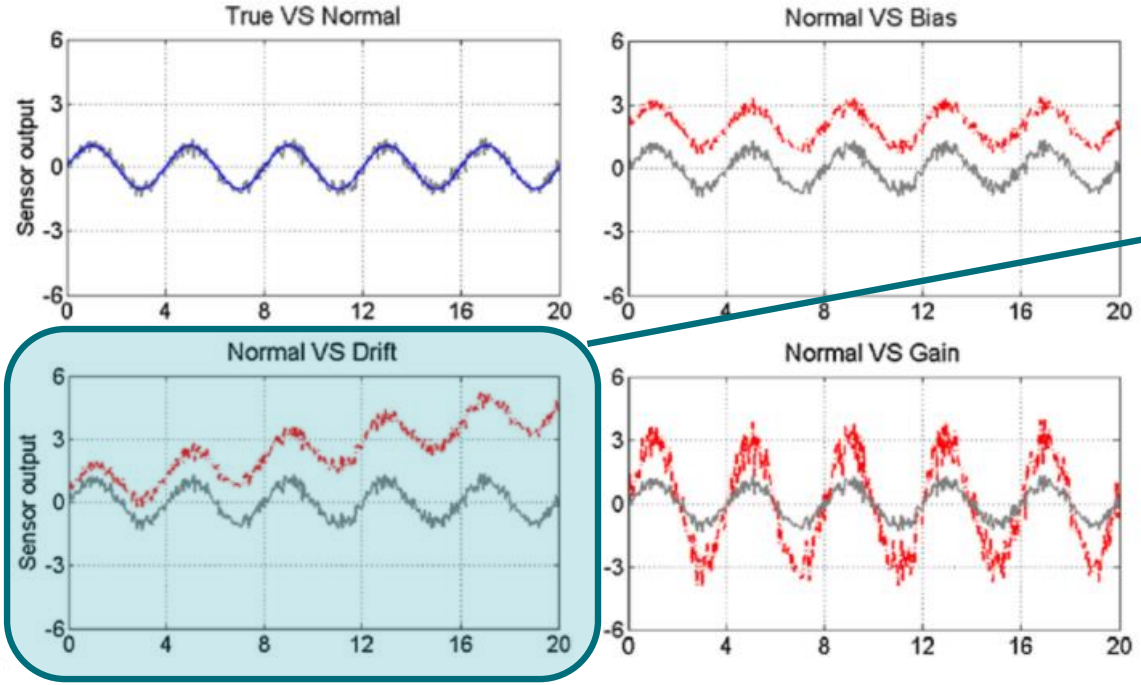
Estudio del efecto de la deriva de sensores para gases en modelos de ML

Autor: Daniel GARCIA TEBA
Tutor: Miguel Ángel GÓMEZ LOPEZ

Estructura

1. Presentación del problema
2. Obtención y procesado de los datos
3. Diseño e implementación de los modelos
4. Conclusiones y planes de mejora

Presentación del problema



Sensor drift es el fenómeno que se produce cuando la diferencia entre el valor esperado y el obtenido varía linealmente con el tiempo.

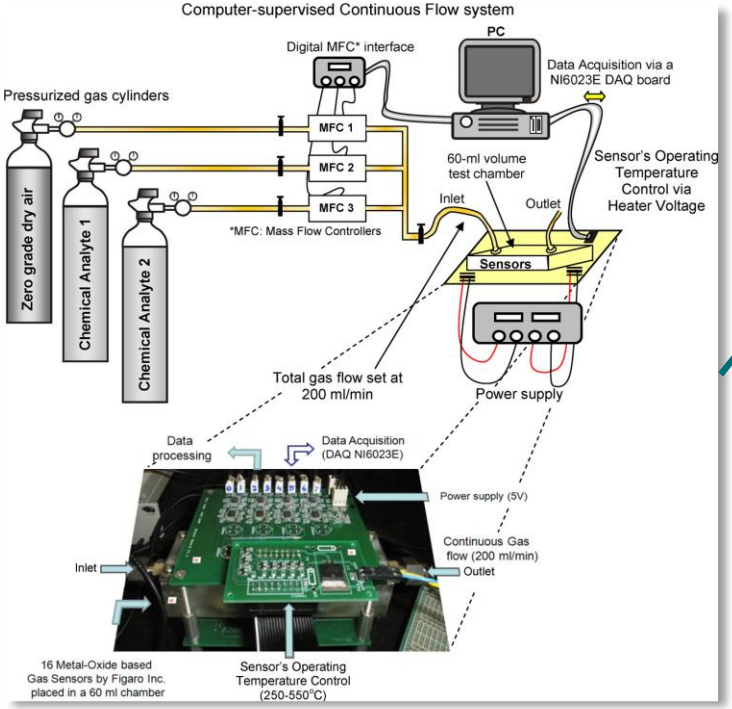
Ref: Yi, Ting-Hua & Huang, Hai-Bin & Li, Hong-Nan. (2017). Development of sensor validation methodologies for structural health monitoring: A comprehensive review. Measurement. 109. 10.1016/j.measurement.2017.05.064.

Obtención y procesado de los datos

Gas Sensor Array Drift Dataset Data Set
https://archive.ics.uci.edu/ml//datasets/Gas+Sensor+Array+Drift+Dataset

Source:
Creators: Alexander Vergara (vergara '@' ucsd.edu)
BioCircutis Institute
University of California San Diego
San Diego, California, USA
Donors of the Dataset: Alexander Vergara (vergara '@' ucsd.edu)
Ramon Huerta (rhuerta '@' ucsd.edu)

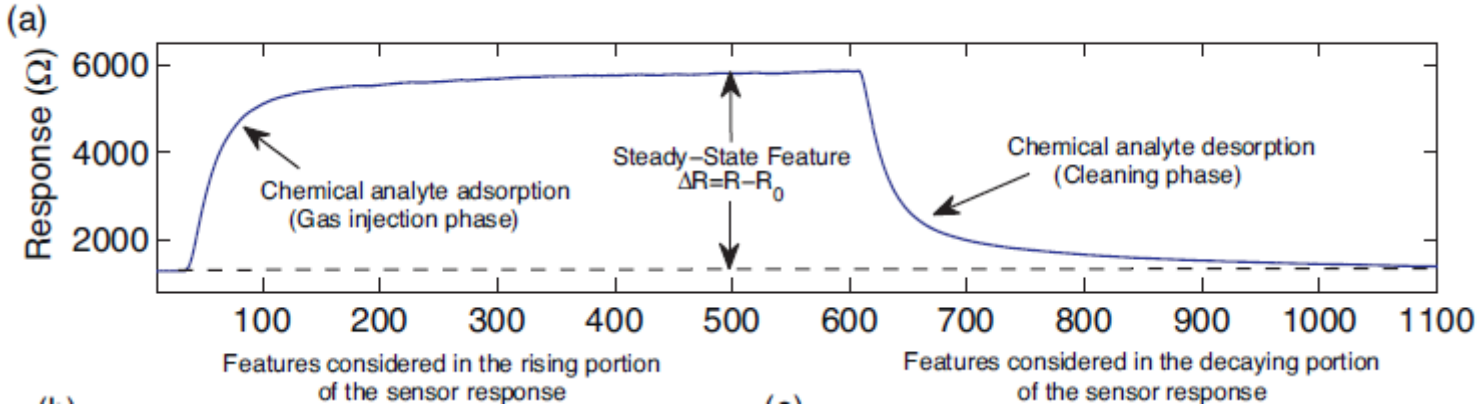
Obtención y procesado de los datos



Esquema del sistema de adquisición, donde pueden verse los 16 sensores de gas Figaro

.

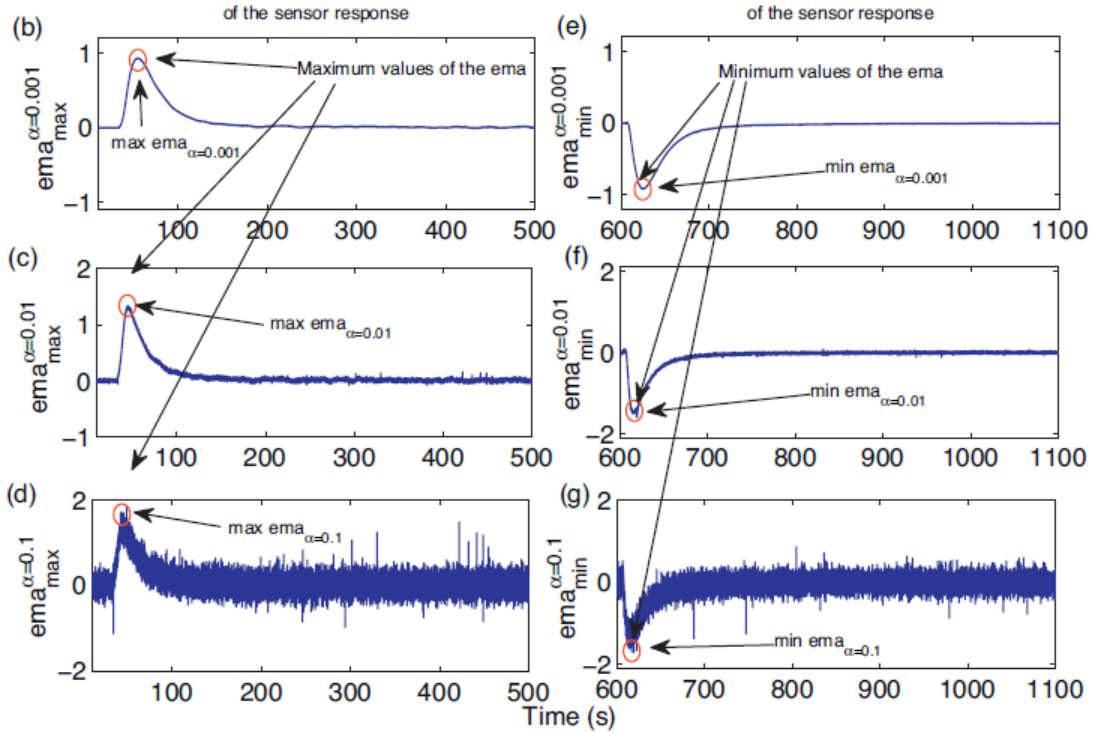
Obtención y procesado de los datos



Steady-State features	Transient features	
	rising portion	decaying portion
ΔR	$MAXema_{\alpha=0.001}$	$MINema_{\alpha=0.001}$
$\|\|\Delta R\|\|$	$MAXema_{\alpha=0.01}$	$MINema_{\alpha=0.01}$
	$MAXema_{\alpha=0.1}$	$MINema_{\alpha=0.1}$

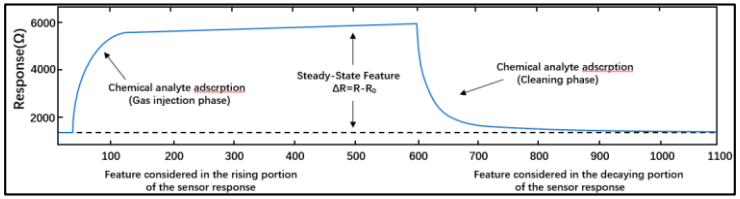
Exponential moving average (EMA) Ref (Vergara y cols., 2011)

Obtención y procesado de los datos



Exponential moving average (EMA) Ref (Vergara y cols., 2011)

Obtención y procesado de los datos



8 features para cada medición de par Gas-Concentración.

16 sensores disponibles

128 features para cada medición de par Gas-Concentración.

¿Qué datos tenemos?

Obtención y procesado de los datos

Batch ID	Month IDs
Batch 1	Months 1 and 2
Batch 2	Months 3, 4, 8, 9 and 10
Batch 3	Months 11, 12, and 13
Batch 4	Months 14 and 15
Batch 5	Month 16
Batch 6	Months 17, 18, 19, and 20
Batch 7	Month 21
Batch 8	Months 22 and 23
Batch 9	Months 24 and 30
Batch 10	Month 36

Tabla 3.1 : Distribución de los lotes a lo largo del tiempo.

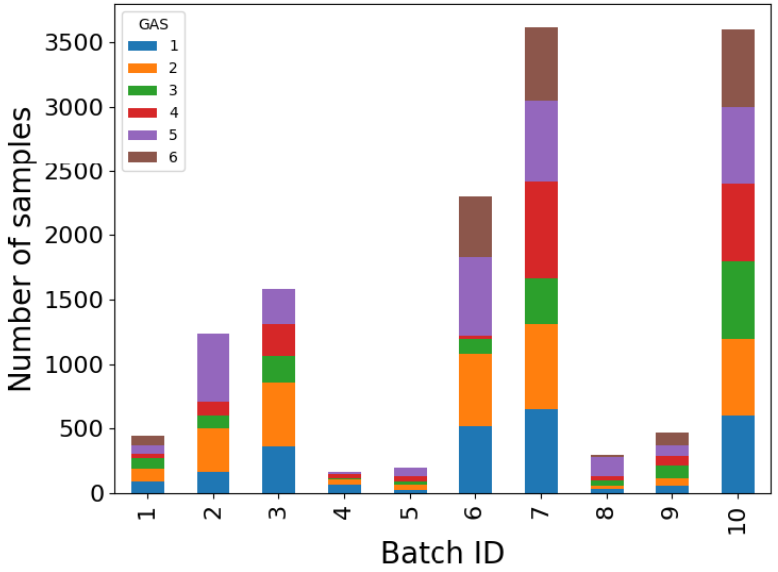
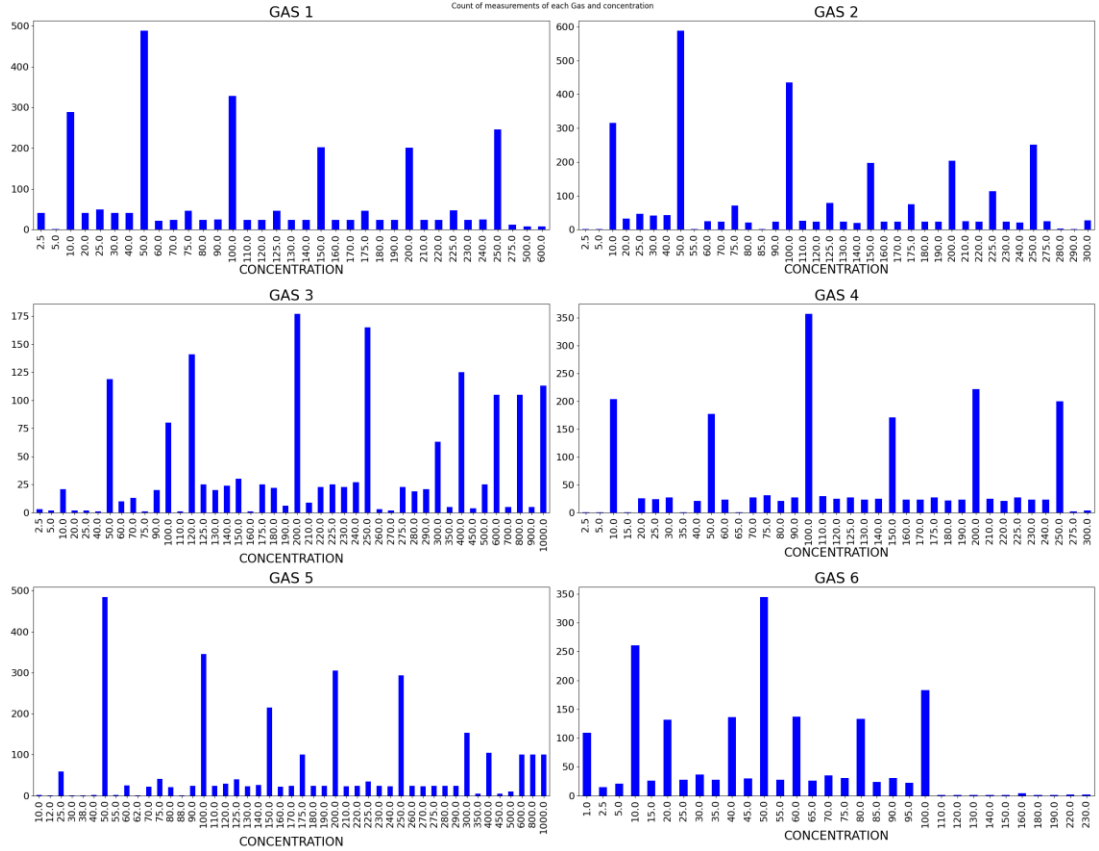


Figura 3.1: Número de muestras de gas por Batch. El número de muestras ensayadas en cada lote es muy desigual, donde los lotes 1,4,5,8 y 9 cuentan con muchas menos mediciones que el resto.

Obtención y procesado de los datos



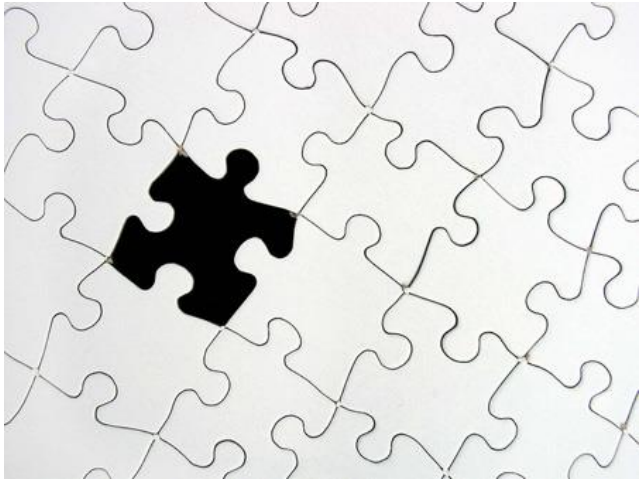
GAS	CONCENTRACIÓN	Numero de muestras
1	50	488
2	50	588
3	200	177
4	100	357
5	50	485
6	50	345

Tabla 3.4. Pares de Gas-Concentración más abundantes

¿Son adecuados para entrenar modelos ML?

Obtención y procesado de los datos

Datos desbalanceados
Diferente número de muestras de cada gas por lote



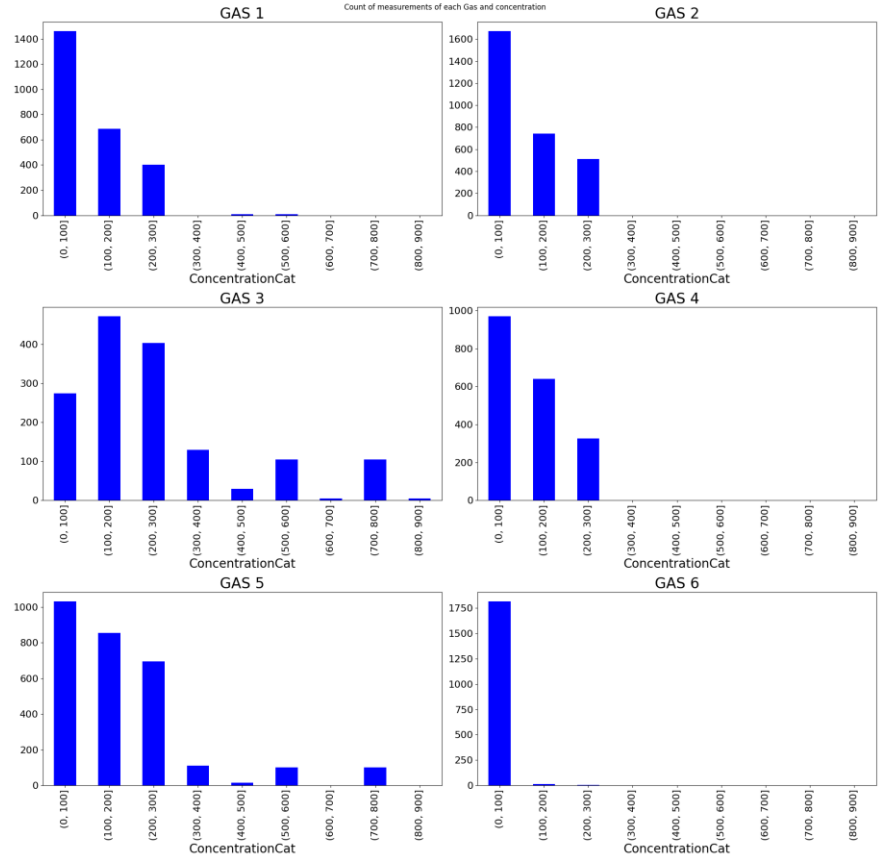
Missing info
En los primeros lotes no hay muestras de gas6

Problema añadido
La concentración de las mediciones de cada gas no son constantes.

Toma aire

Photo by Simon Migaj on Unsplash

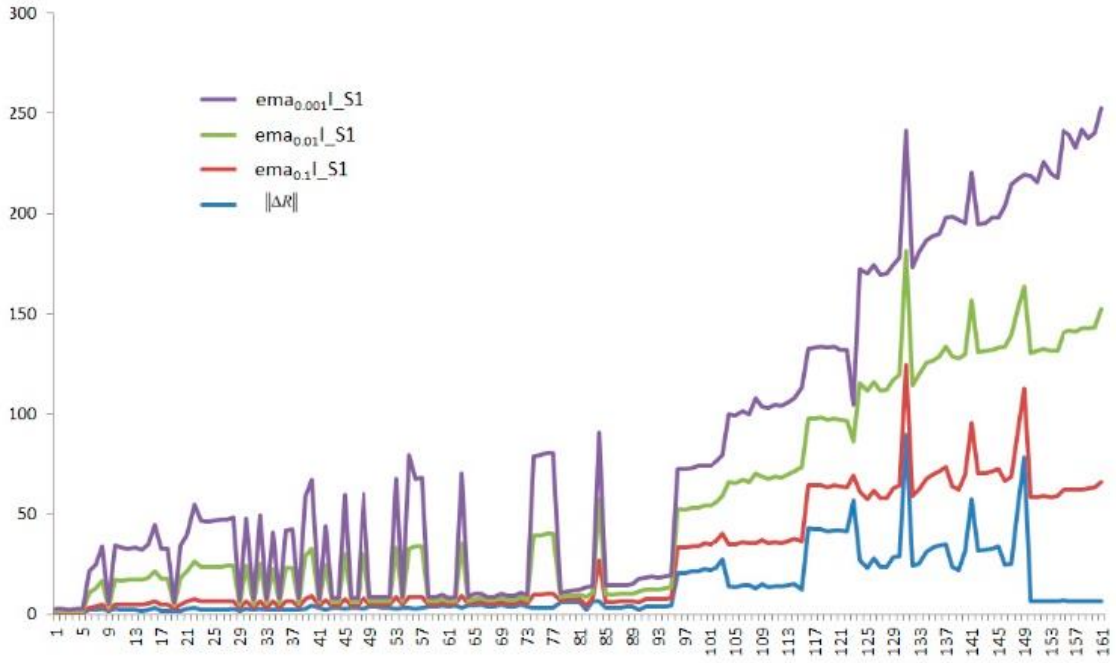
Obtención y procesado de los datos



GAS	CONCENTRACIÓN	Numero de muestras
1	50	488
2	50	588
3	200	177
4	100	357
5	50	485
6	50	345

Tabla 3.4. Pares de Gas-Concentración más abundantes

Obtención y procesado de los datos



Ref Zhao, 2019

Las variables generadas por la descomposición de la señal, están **correlacionadas**.

Obtención y procesado de los datos

Table 1. Sensors Information in the Sensor Array.

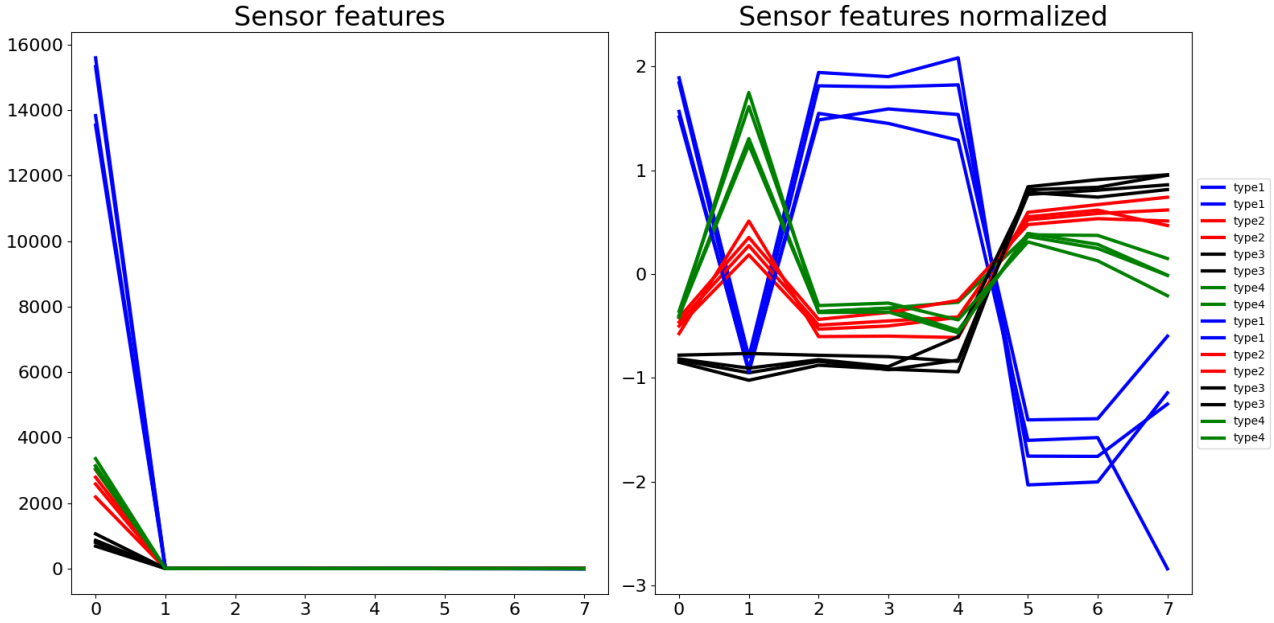
Sensor Type	Number of Units	Target Gases
TGS2600	4	Hydrogen, carbon, monoxide
TGS2602	4	Ammonia, H_2S, volatile organic compounds (VOC)
TGS2610	4	Propane
TGS2620	4	Carbon monoxide, combustible gases, VOC

Además, los **16 sensores** han sido calibrados con diferentes sensibilidades, y cada tipo de sensor es más eficiente para detectar un gas concreto.

Esto significa que

- Habrá mediciones que **saturen** algunos sensores, y otros sensores no sean capaz de **detectar nada**.
- Las mediciones entre sensores estarán **correladas**.

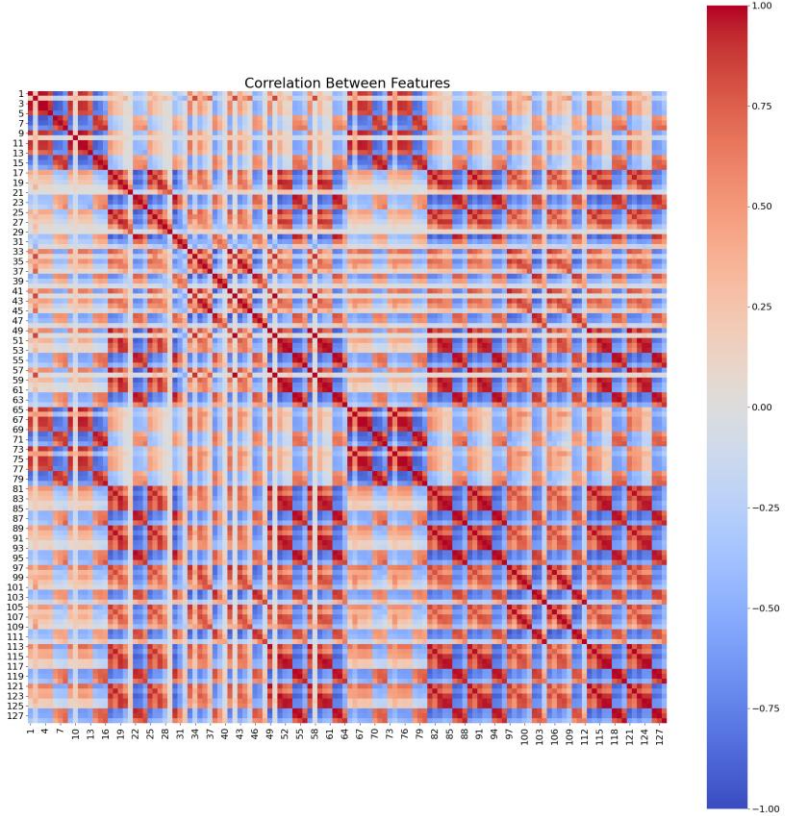
Obtención y procesado de los datos



En el Eje X están las 8 features extraidas para el gas, y en el ejeY su valor.

Sensores del mismo tipo tienden a generar señales muy similares.

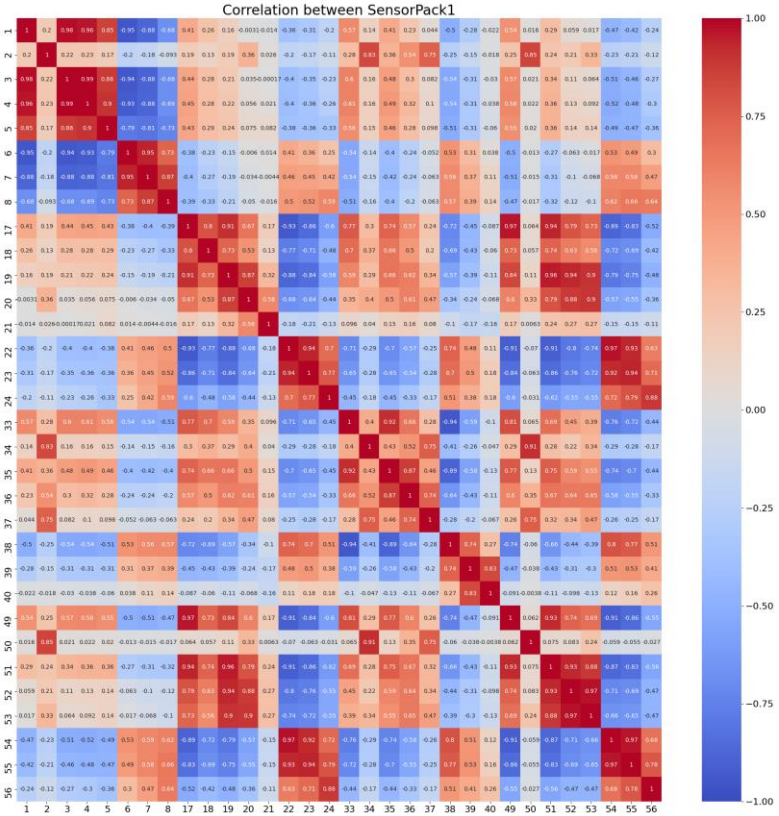
Obtención y procesado de los datos



Se ha calculado la matriz de correlación para las 8 componentes de los 16 sensores (128).

Puede observarse cómo para un mismo sensor la correlación entre features es fuerte, y la correlación entre sensores también existe.

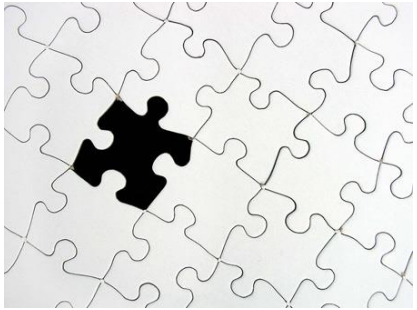
Obtención y procesado de los datos



Si escogemos un sensor de cada tipo, y calculamos de matriz de confusión, vemos que la **correlación sigue siendo fuerte.**

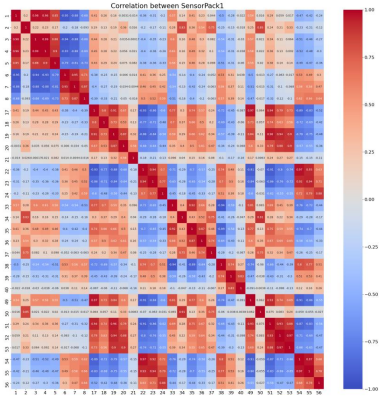
Obtención y procesado de los datos

Datos desbalanceados
Diferente número de muestras de cada gas por lote



Missing info
En los primeros lotes no hay muestras de gas6

Datos correlacionados
Entre sensores y entre features



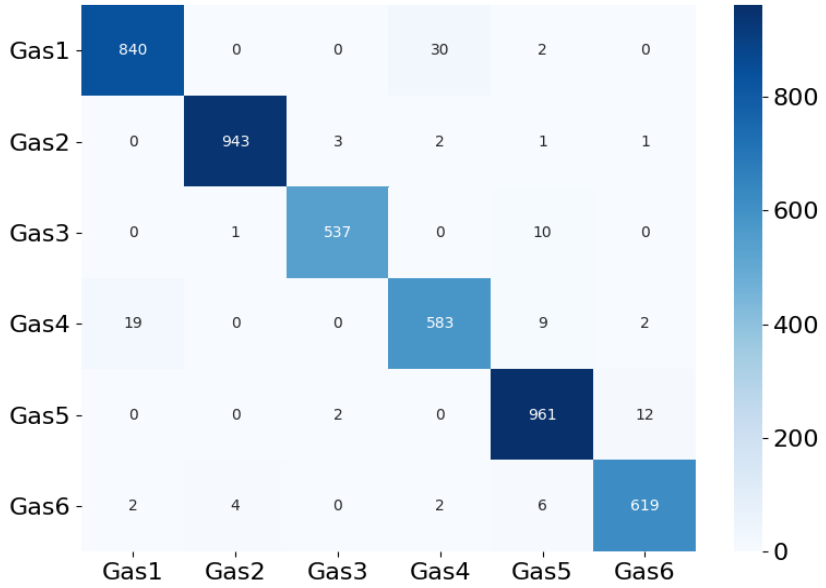
Eliminado la variable concentración del problema

Photo by Braden Collum on Unsplash

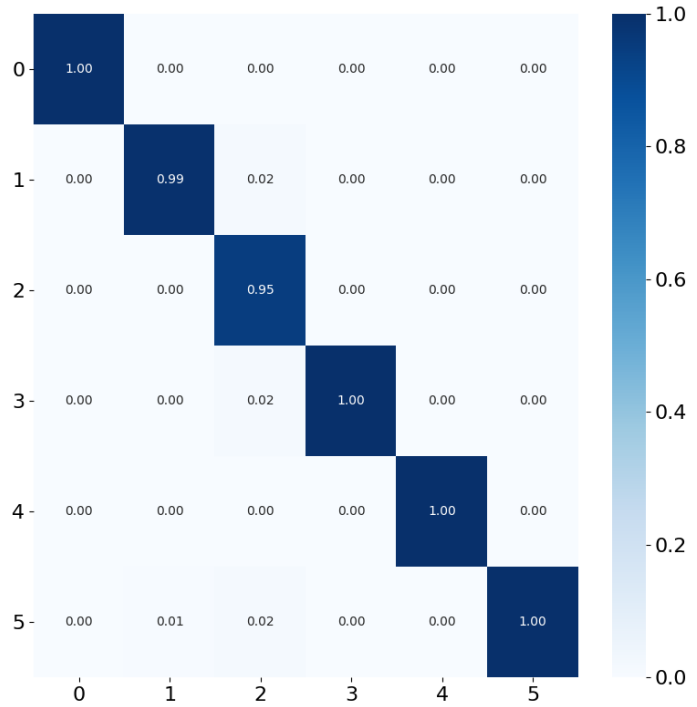
Listos…

Diseño e implementación de los modelos

Modelos supervisados. Modelo de red neuronal secuencial



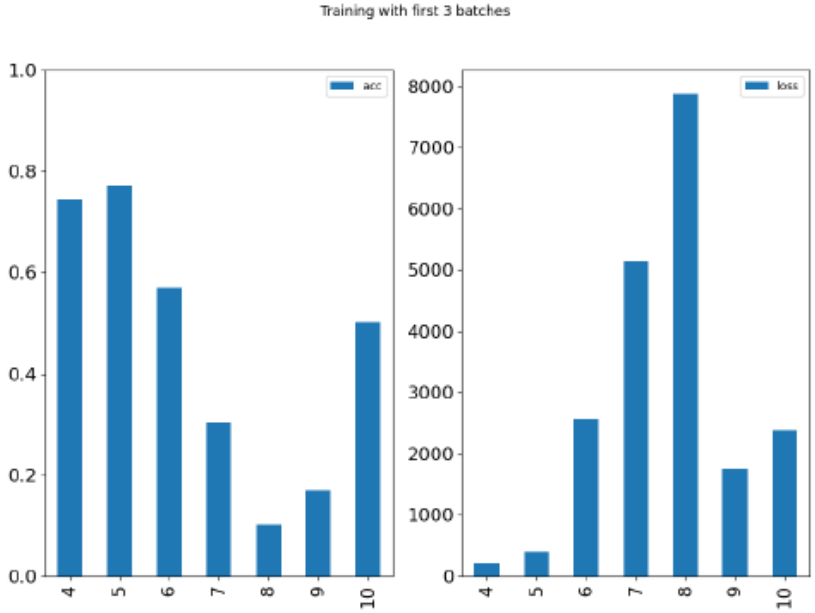
Matriz de confusión obtenida con train-test Split al 70/30. Valores absolutos.

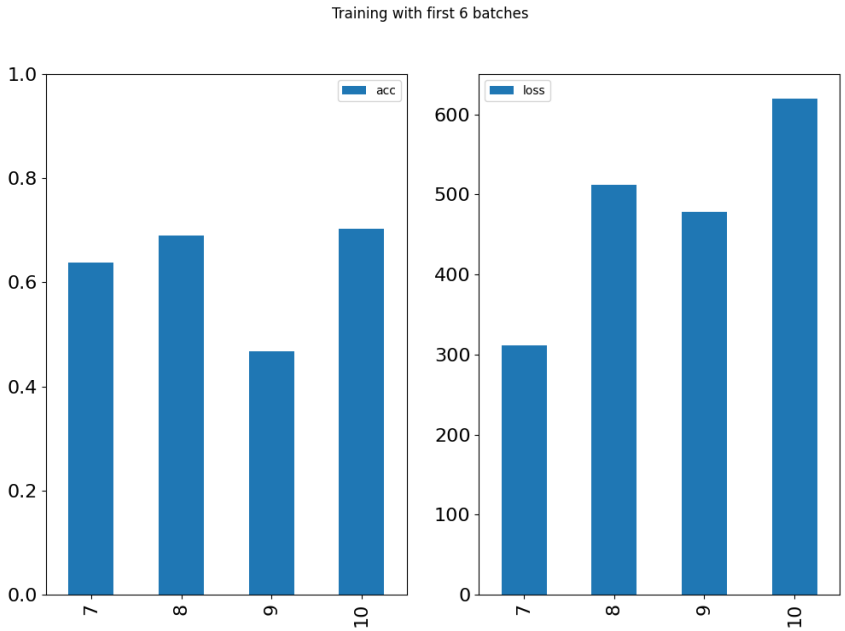


Matriz de confusión obtenida usando lotes 1 al 9 de entrenamiento y lote 10 para la validación. Valores relativos.

Diseño e implementación de los modelos

Modelos supervisados. Modelo de red neuronal secuencial



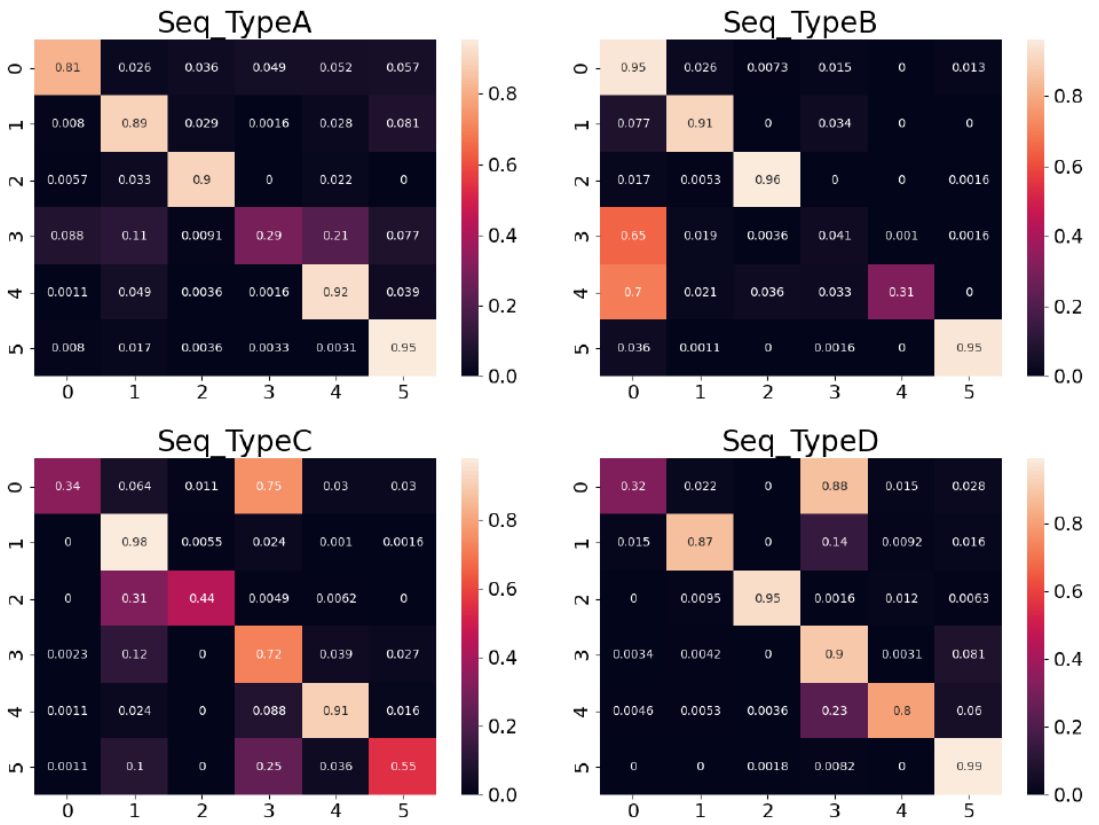


Diseño e implementación de los modelos

Modelo de red neuronal secuencial

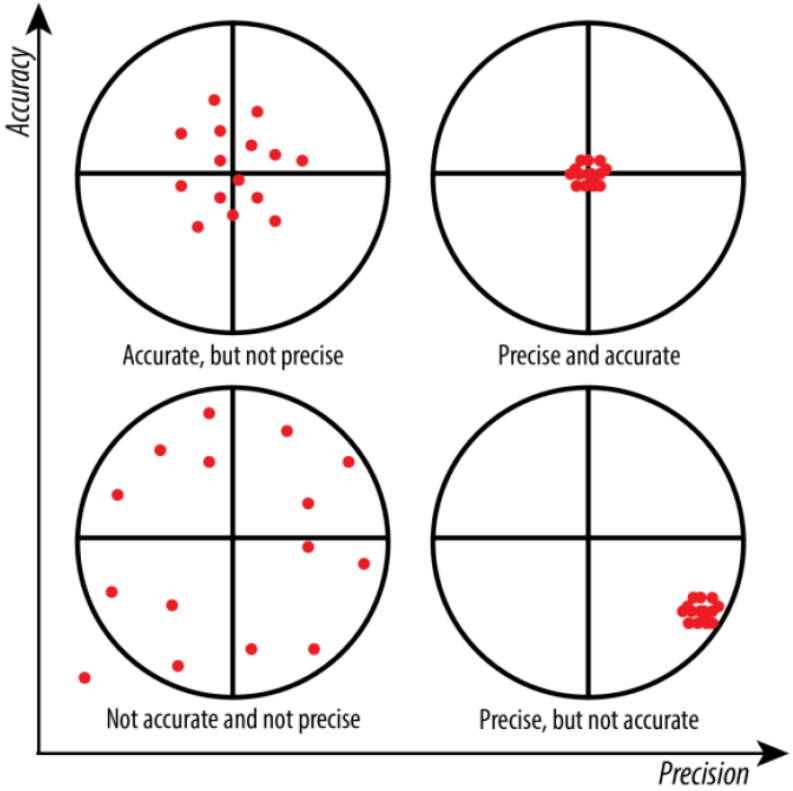
Figura: Matriz de confusión utilizando solo sensores **tipo A** en la img sup izq, solo sensores **tipo B** en la img sup derch y así sucesivamente.

- El sensor tipo A predice bastante bien **todos los tipos** de gases
- los sensores B **no consiguen** detectar dos gases
- Los tipo C y D **fallan** mucho al detectar un gas en concreto.



Diseño e implementación de los modelos

Extra



Diseño e implementación de los modelos

Modelos no supervisados. PCA + KMeans

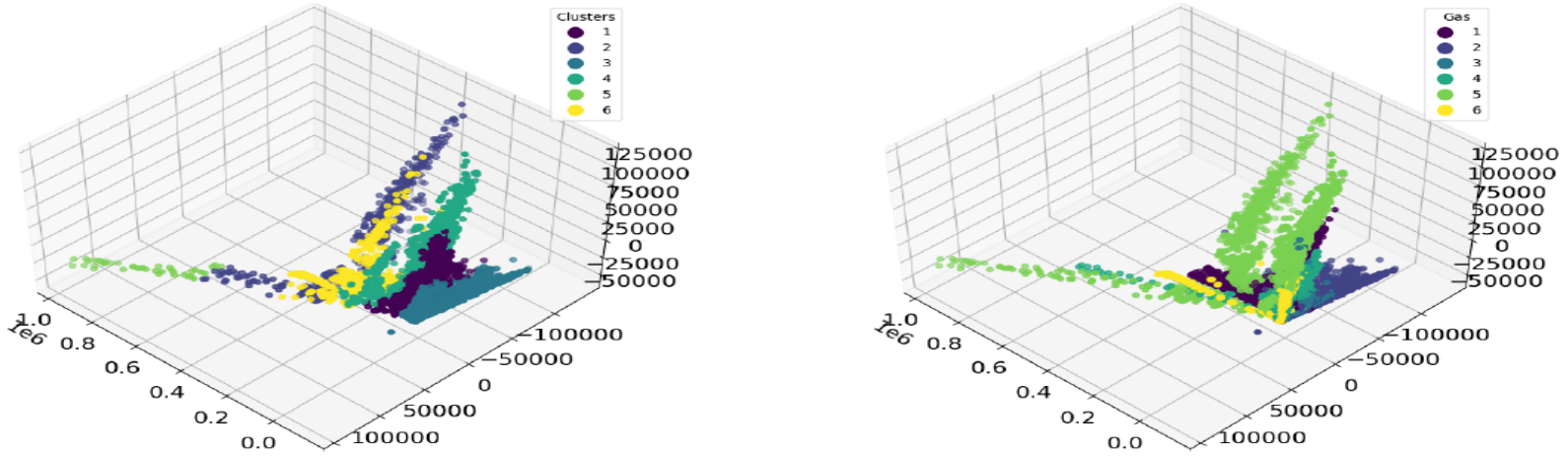
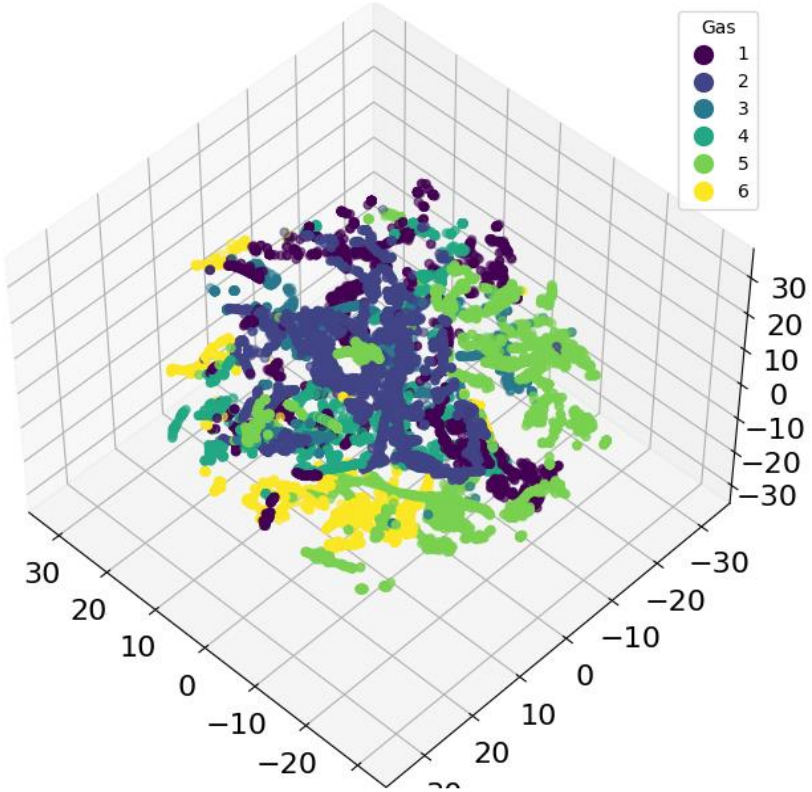


Figura.
A la izquierda, color según clusters detectados.
A la derecha, color según el gas al que pertenece.

Diseño e implementación de los modelos

Modelos no supervisados. TSNE

Figura.
Color según el gas al que pertenece.



Diseño e implementación de los modelos

Modelos no supervisados. PCA+ KMeans

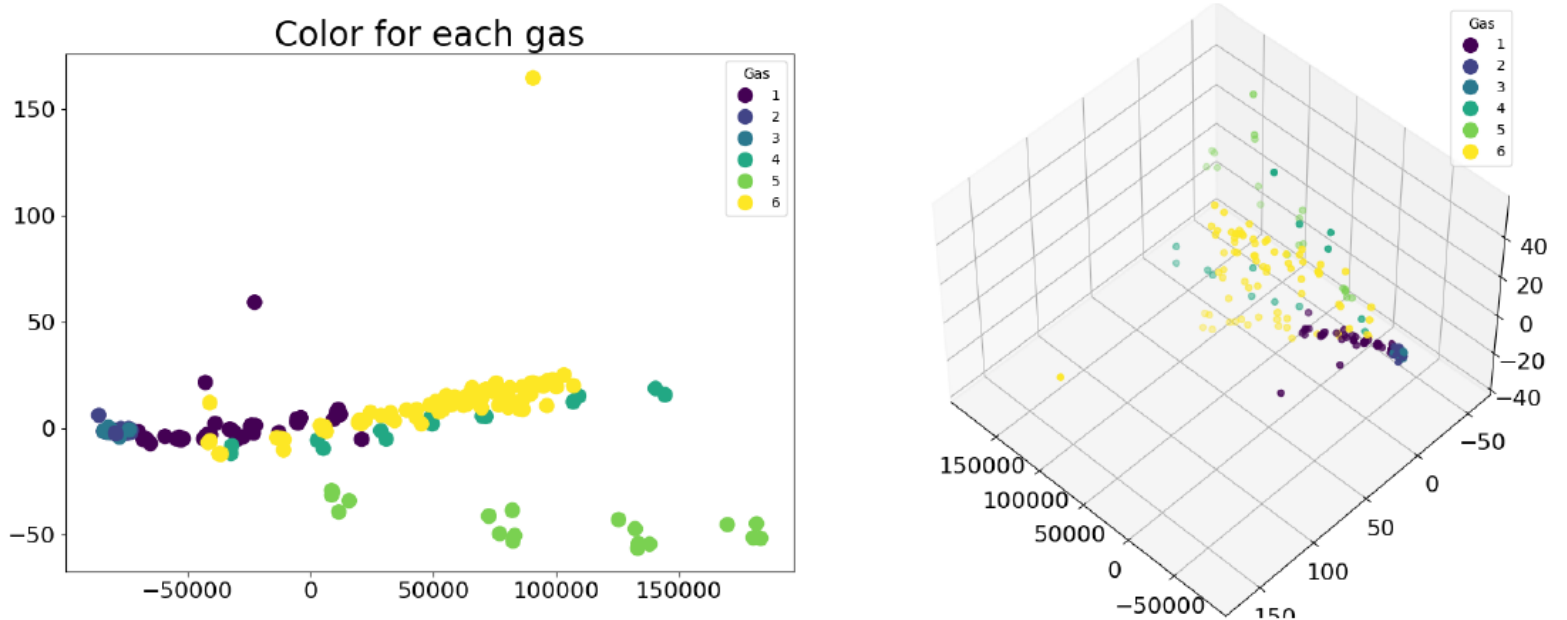


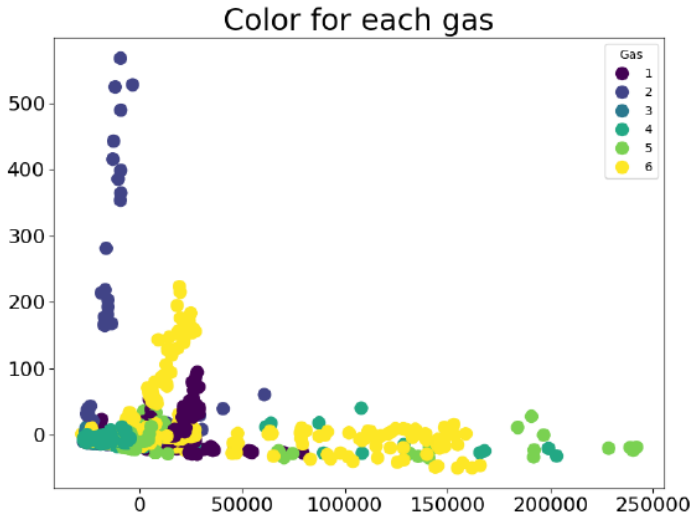
Figura.

Resultados PCA para los datos del batch 1, sensor1 y concentraciones por debajo de 100ppmv. Se han coloreado los puntos según a qué gas pertenece.

A la izq se ha reducido la dimensionalidad a 2d, y a la derecha a 3d. Los gases aparecen en clusters bien diferenciados.

Diseño e implementación de los modelos

Modelos no supervisados. PCA+ KMeans



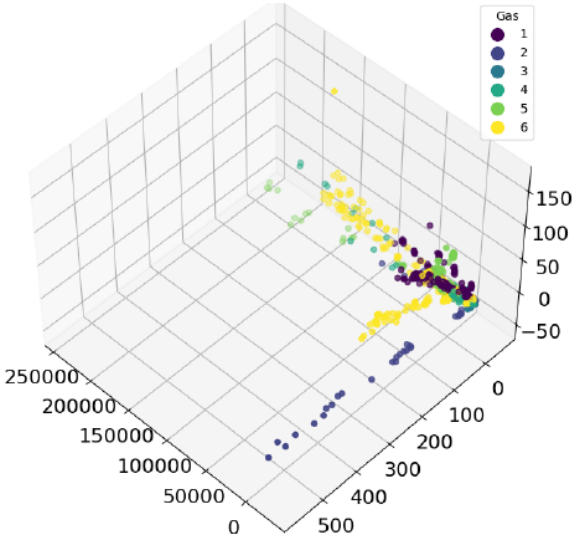


Figura.

Resultados PCA para los datos del batch 1 y 10, sensor1 y concentraciones por debajo de 100ppmv. Se han coloreado los puntos según a qué gas pertenece.

Los puntos que han aparecido con respecto a la imagen anterior no se han agrupado con el cluster del batch 10, si no que han generado una nueva rama, lo que nos indica que son lo suficiente diferentes como para estar agrupadas aparte.

Conclusiones y planes de mejora

- La deriva o drift en los sensores tiene un **efecto muy negativo en la capacidad de predicción de los modelos de regresión,** que no puede obviarse. Tanto redes neuronales, randomForest o LightGBM.

- Los métodos de clasificación basados en redes neuronales **son lentos en entrenar**, y su **accuracy** se va reduciendo conforme nos alejamos de los datos de entrenamiento.

- Los métodos basados en RandomForest o LightGBM entrenan con mucha rapidez, pero **no son inmunes al efecto** del drift y el **accuracy** desciende con mediciones distantes entre sí.

- Todos los métodos anteriormente mencionados **fallan en precisión** a causa del drift, **no en accuracy**, catalogando unos gases de forma errónea en la categoría de otro gas.

- Los métodos de aprendizaje no supervisado Kmeans y TSNE **tienen un rendimiento mucho peor que las redes neuronales o los RF o LGBM**. TSNE es muy costoso computacionalmente.

Gracias por su atención

¿Preguntas?

Siempre son bienvenidas

Referencias

[Ref1] Vergara, A., Ayhan, T., Vembu, S., Huerta, R., Ryan, M., y Homer, M. (2011,01). Gas sensor drift mitigation using classifier ensembles.
doi: 10.1145/2003653.2003655

[Ref2] Vergara, A., Vembu, S., Ayhan, T., Ryan, M. A., Homer, M. L., y Huerta, R.(2012). Chemical gas sensor drift compensation using classifier ensembles.
Sensors and Actuators, B: Chemical, *166-167*, 320–329.
doi: 10.1016/j.snb.2012.01.074

[Ref3] Zhao, H., Li, L., Xiao, W., Meng, Z., Han, y Yu, H. (2019, 09).
Sensor drift compensation based on the improved lstm and svm multi-class ensemble learning models. *Sensors*, *19*, 3844. doi: 10.3390/s19183844