TRABAJO FIN DE MÁSTER

Master Universitario en Gestión y Análisis de Grandes Volúmenes de Datos: BIG DATA

CURSO ACADÉMICO 2019-2020

UEMC Universidad Europea Miguel de Cervantes

Escuela de Negocios y Dirección

I. CONSIDERACIONES GENERALES

II. CALENDARIO DE TFM

I. Consideraciones generales

1. Documento de referencia (en "materiales de referencia" del campus):

https://www.opencampus.uemc.es/pluginfile.php/251601/mod_resource/content/34/3_Directrices%20para%20el%20desarrollo%20del%20TFM_MBIGDATA.pdf

El Trabajo de Fin de Máster (TFM) supone la realización por parte del estudiante de un proyecto, memoria o trabajo original, autónomo e individual que se llevará a cabo bajo la orientación de un tutor y que permitirá al estudiante mostrar de forma integrada los contenidos formativos recibidos y las capacidades, competencias y habilidades adquiridas asociadas al título de Máster.

El estudiante deberá desarrollar el TFM y presentarlo para su defensa oral ante un tribunal de tal manera que demuestre los conocimientos y capacidades adquiridos en las áreas de conocimiento del correspondiente Máster.

I. Consideraciones generales

2. Planificación

- Un primer video de presentación de la asignatura y de la normativa que le aplica.
- Una segunda tutoría que será impartida por cada tutor a sus alumnos en los que explicará la dinámica de trabajo, entregas, plazos y condiciones para superar la asignatura.
- Actividad 1. En la que cada estudiante debe cumplimentar una encuesta en la que elija el caso sobre el que realizará el TFM, así como la convocatoria en la que quiere desarrollarlo.
- Actividad 2. Entrega parcial del TFM, en la que el tutor le dará seguimiento de los avances del trabajo y podrá contestar dudas del alumno.
- Actividad 3. Entrega final del TFM, en la que el tutor debe emitir un informe dando APTO o NO APTO al TFM para que pueda ser solicitada la defensa ante tribunal

I. Consideraciones generales

2. Memoria

- **Extensión: 40-50 páginas (sin anexos)**
- **Formato:** https://www.opencampus.uemc.es/mod/resource/view.php?id=189233
- **Normas APA:** https://www.opencampus.uemc.es/pluginfile.php/251604/mod_resource/content/33/6_Normas%20APA_UEMC.pdf
- **Estructura orientativa:**
  - Objetivos del trabajo.
  - Análisis de la situación.
  - Obtención, procesado y almacenamiento de los datos.
  - Análisis exploratorio.
  - Diseño e implementación de los modelos o técnicas necesarios.
  - Análisis de los resultados obtenidos.
  - Conclusiones y planes de mejora.
  - Bibliografía.
  - Anexo con el código fuente desarrollado.

I. Consideraciones generales

2. Evaluación

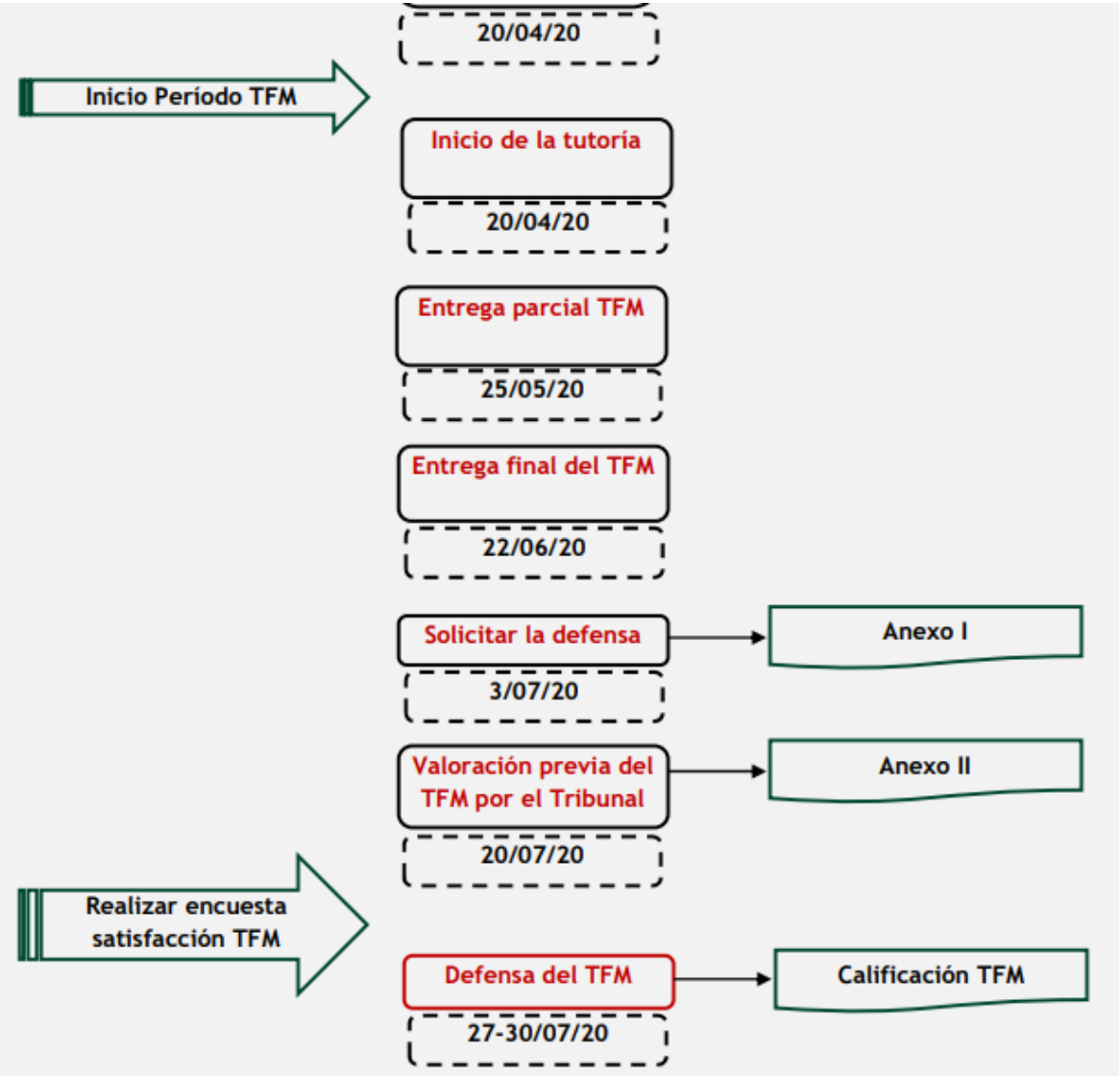
- **Memoria: evaluada por el tutor**
- **Defensa: tribunal de dos o tres personas (sin el tutor)**
  - **Por videoconferencia.**
  - **Aproximadamente 10 min de presentación y 5 de preguntas.**

Título	% EF
Pruebas orales (Tribunal)	20%
Trabajos y proyectos (Tutor 50% y Tribunal 20%)	70%
Escala de actitudes (Tutor)	10%

I. CONSIDERACIONES GENERALES

II. CALENDARIO DE TFM

1. Calendario de TFM: Etapas del proceso – Convocatoria ordinaria



2. Calendario de TFM: Etapas del proceso – Convocatoria extraordinaria

https://www.opencampus.uemc.es/pluginfile.php/251600/mod_resource/content/35/2_Calendario%20TFM_M_BIGDATA_MBA.pdf

Apertura del aula: Elección caso y elección de periodo de tutorización (misma encuesta)
Videoconferencia de Directores a los alumnos para presentar la asignatura
Fin de plazo para elegir caso y periodo. Se volverá a abrir el plazo de elección de caso de nuevo entre el 1-15 de julio para los que no hayan contestado y quieran realizar el TFM en la convocatoria 2
Fecha límite de asignación de matriculación de alumnos al TFM (para alumnos que se matriculen despues de las actas del primer semestre)
Inicio TFM y tutorización a los alumnos
Entrega parcial y final del TFM por parte del alumno
Corrección y conformidad del tutor
Periodo para que los alumnos soliciten la defensa condicionado a que tengan la conformidad del tutor
Cierre actas bloque 4. Fecha limite para tener superadas las PE
Validación de solicitudes
Selección de tribunales y publicación de fechas provisionales para la defensa
Envio de los TFM a los tribunales y periodo para que realicen su informe
Publicación en el campus de los informes de los tribunales y fechas definitivas para la defensa de los TFM
Realización de las defensas de TFM y calificación

MAYO

L	M	X	J	V	S	D
				1	2	3
4	5	6	7	8	9	10
11	12	13	14	15	16	17
18	19	20	21	22	23	24
25	26	27	28	29	30	31

JUNIO

L	M	X	J	V	S	D
1	2	3	4	5	6	7
8	9	10	11	12	13	14
15	16	17	18	19	20	21
22	23	24	25	26	27	28
29	30					

JULIO

L	M	X	J	V	S	D
29	30	1	2	3	4	5
6	7	8	9	10	11	12
13	14	15	16	17	18	19
20	21	22	23	24	25	26
27	28	29	30	31	1	2

AGOSTO

L	M	X	J	V	S	D
					1	2
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28	29	30
31						

Convocatoria 2

SEPTIEMBRE

L	M	X	J	V	S	D
	1	2	3	4	5	6
7	8	9	10	11	12	13
14	15	16	17	18	19	20
21	22	23	24	25	26	27
28	29	30				

6
7

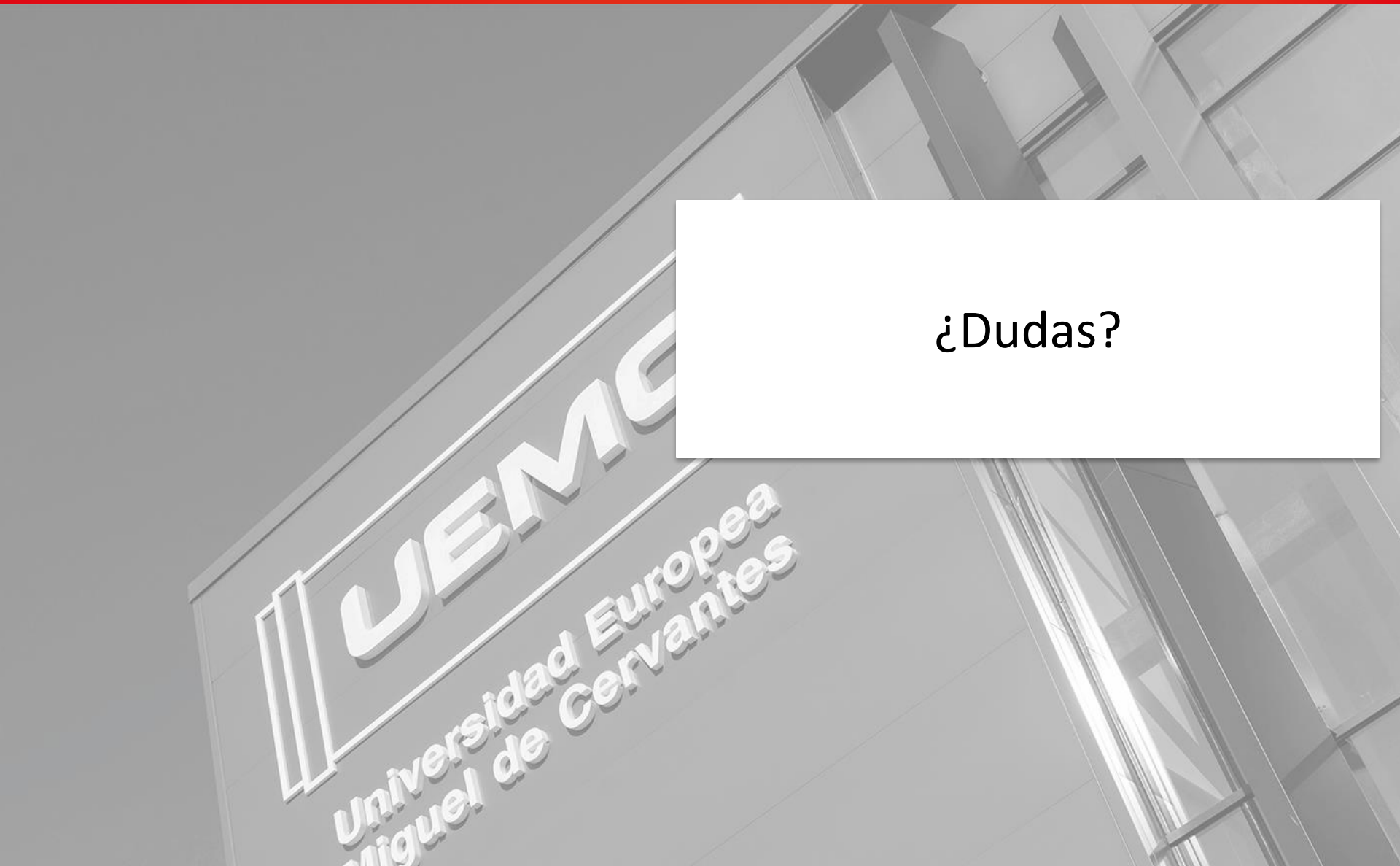
OCTUBRE

L	M	X	J	V	S	D
			1	2	3	4
5	6	7	8	9	10	11
12	13	14	15	16	17	18
19	20	21	22	23	24	25
26	27	28	29	30	31	

NOVIEMBRE

L	M	X	J	V	S	D
						1
2	3	4	5	6	7	8
9	10	11	12	13	14	15
16	17	18	19	20	21	22
23	24	25	26	27	28	29
30						

UEMC universidad online



¿Dudas?

- Situación y antecedentes
- Objetivos del trabajo
- Descripción del problema
- Análisis de la situación
- Solución propuesta
- Desarrollo de la solución
- Conclusiones y planes de mejora
- Bibliografía
- Anexos, ej:
  - Manual de usuario
  - Manual de instalación
  - …
- Entrega de todo el software realizado, debidamente documentado

- En este trabajo se plantea el análisis de una serie de datos experimentales obtenidos de una nariz electrónica, la cual será usada para la adquisición y discriminación de distintos odorantes bajo estudio, en particular, todo tipo de contaminantes medioambientales y sustancias peligrosas.
- Se plantea el uso de dataset disponibles de forma libre (10mb de dataset).
- Es posible explorar diversas aproximaciones al problema, comparando resultados:
  - Algoritmos “clásicos”: Inferencia Bayesiana, Random Forest, SVM
  - Redes diseñadas desde cero, o reutilizando ya existentes
- Sería muy interesante proseguir hacia el problema de detección con bajo nivel de falsas alarmas (error Tipo I, o alfa, falso positivo)

Datos

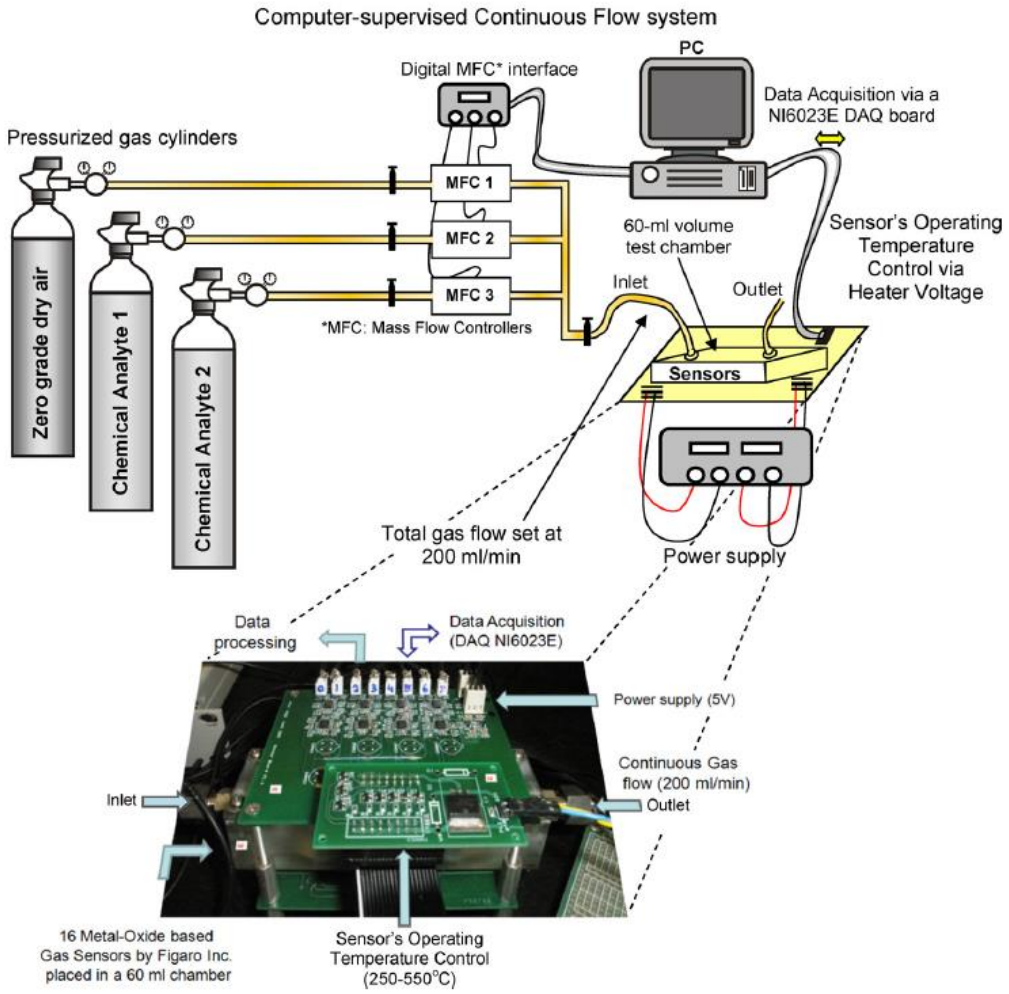


Fig. 1. Experimental setup used for data acquisition. The sensor responses are recorded in the presence of the analyte in gaseous form diluted at different concentrations in dry air. The measurement system operates in a fully computerized environment with minimal human intervention, which provides versatility in conveying the odors of interest (at the desired concentrations) to the sensing chamber with high accuracy, and simultaneously in keeping constant the total flow. Therefore, no changes in the flow or flow dynamics are reflected in the sensor response (i.e., only the presence of an odorant will be reflected in the sensor response). Moreover, since the system continuously supplies gas to the sensing chamber (either clean dry air or a chemical component), the amount of gas molecules in the sensing chamber is homogeneously distributed.

Configuración experimental utilizada para la adquisición de datos. Las respuestas del sensor se registran en presencia del analito en forma gaseosa diluida a diferentes concentraciones. en aire seco El sistema de medición funciona en un entorno totalmente computarizado con mínima intervención humana, lo que proporciona versatilidad para transmitir los olores de interés (a las concentraciones deseadas) a la cámara de detección con alta precisión, y simultáneamente en mantener constante el flujo total. Por lo tanto, no hay cambios en el flujo. o la dinámica del flujo se refleja en la respuesta del sensor (es decir, solo la presencia de un olor se reflejará en la respuesta del sensor). Además, ya que el sistema continuamente suministra gas a la cámara de detección (ya sea aire limpio y seco o un componente químico), la cantidad de moléculas de gas en la cámara de detección se distribuye de manera homogénea.

Datos

Las respuestas de dichos sensores se leen en forma de resistencia a través de la capa activa de cada sensor; por lo tanto, cada medición produjo una serie temporal de 16 canales, cada una representada por un conjunto de características que reflejan los procesos dinámicos que ocurren en la superficie del sensor en reacción a la sustancia química que se está evaluando.

http://archive.ics.uci.edu/ml/datasets/Gas%2BSensor%2BArray%2BDrift%2BDataset%2Bat%2BDifferent%2BConcentrations

Datos

En particular, se consideraron dos tipos distintos de características en la creación de este conjunto de datos:

- la llamada característica de **estado estable (DR)**, definida como el cambio de resistencia máxima con respecto a la línea base
  - versión **DR** normalizada (**DR** dividido por el valor adquirido cuando el vapor químico está presente en la cámara de prueba).
- un conjunto de características que reflejan la dinámica del sensor de la porción transitoria creciente / decreciente de la respuesta del sensor durante toda la medición.
  - Este conjunto de características es una transformación, tomada del campo de la econometría y originalmente presentada a la comunidad de quimio-detección por Muezzinoglu et al. (2009).
  - La parte transitoria de la respuesta del sensor se convierte en un escalar real mediante la estimación del valor máximo / mínimo y[k] para la porción ascendente / decreciente de la media móvil exponencial de la respuesta del sensor:

$$y[k] = (1-Alfa)\, y[k-1]+Alfa(R[k]-R[k-1])$$

donde **R [k]** es la resistencia del sensor medida en el tiempo k y Alfa es un parámetro de suavizado escalar entre 0 y 1.

En particular, se establecieron tres valores diferentes para Alfa = 0.1, 0.01, 0.001 para obtener tres valores de características diferentes de la porción ascendente de la respuesta del sensor y tres características adicionales con los mismos valores Alfa para la porción decadente de la respuesta del sensor, cubriendo así toda la dinámica de respuesta del sensor.

Datos

Para fines de procesamiento, el conjunto de datos está organizado en diez lotes, cada uno con el número de mediciones por clase y mes indicado en las tablas a continuación. Esta reorganización de los datos se realizó para garantizar un número de experimentos suficiente y tan uniformemente distribuido como sea posible en cada lote.

ID de lote ID de mes
Lote 1 Meses 1 y 2
Lote 2 meses 3, 4, 8, 9 y 10
Lote 3 meses 11, 12 y 13
Lote 4 meses 14 y 15
Lote 5 Mes 16
Lote 6 meses 17, 18, 19 y 20
Lote 7 Mes 21
Lote 8 meses 22 y 23
Lote 9 meses 24 y 30
Lote 10 Mes 36

ID de lote: etanol, etileno, amoníaco, acetaldehído, acetona, tolueno
Lote 1: 83, 30, 70, 98, 90, 74
Lote 2: 100, 109, 532, 334, 164, 5
Lote 3: 216, 240, 275, 490, 365, 0
Lote 4: 12, 30, 12, 43, 64, 0
Lote 5: 20, 46, 63, 40, 28, 0
Lote 6: 110, 29, 606, 574, 514, 467
Lote 7: 360, 744, 630, 662, 649, 568
Lote 8: 40, 33, 143, 30, 30, 18
Lote 9: 100, 75, 78, 55, 61, 101
Lote 10: 600, 600, 600, 600, 600, 600

Datos

El conjunto de datos está organizado en archivos, cada uno representando un lote diferente. Dentro de los archivos, cada línea representa una medida. El primer carácter (1-6) codifica el producto (analyte), seguido del nivel de concentración:

1: etanol; 2: etileno; 3: amoniaco; 4: acetaldehído; 5: acetona; 6: tolueno

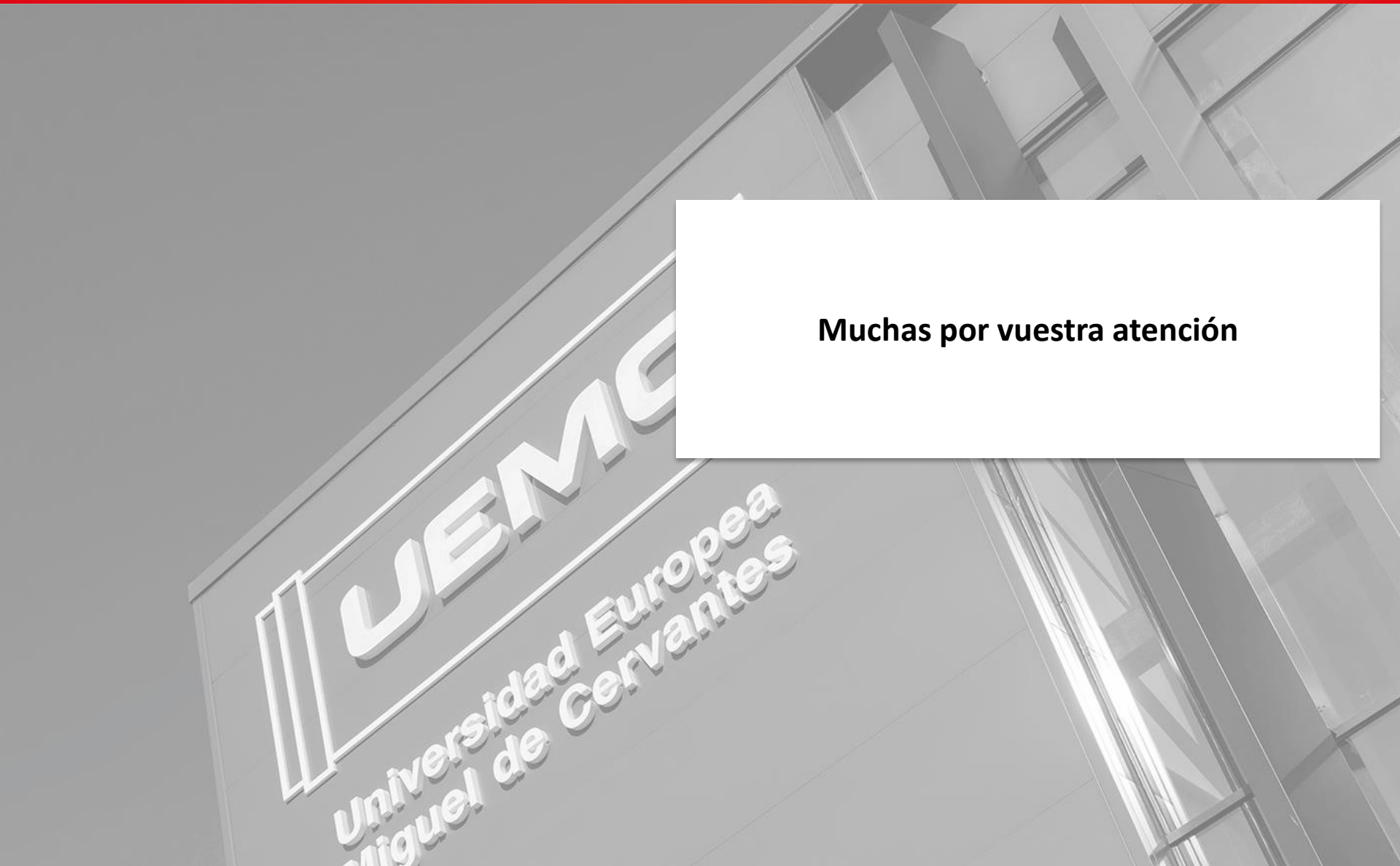
El formato de datos sigue el mismo estilo de codificación que en el formato libsvm x: v, donde x representa el número de característica y v el valor real de la característica. Por ejemplo, en
1; 10.000000 1: 15596.162100 2: 1.868245 3: 2.371604 4: 2.803678 5: 7.512213 128: -2.654529

El número 1 representa el número de clase (en este caso, etanol), el nivel de concentración de gas fue de 10ppmv, y las 128 columnas restantes enumeran los valores de características reales para cada registro de medición organizado como se describió anteriormente.

Recomendaciones

- **El problema básico es sencillo, pero el objetivo es tratar de optimizarlo, ir más allá que en los trabajos de la asignatura.**
- **No perderse en el dataset. Antes de lanzarse a programar gastar un par de semanas en entender bien que se ha medido, como, y que se ha reflejado en el datasheet. Hay mucho trabajo "ya hecho" que se puede aprovechar.**
- **Recomiendo conocer los datos antes de comenzar.**
- **Recomiendo leer las tres referencias del dataset antes de comenzar:**
  - **https://doi.org/10.1016/j.snb.2012.01.074**
  - **https://doi.org/10.1016/j.chemolab.2013.10.012**
  - **10.1016/j.snb.2008.10.065**
- **Son datasets más o menos contenido, la forma habitual de trabajar de clase, cargando todo en RAM, probablemente servirá.**
- **Saber por qué se clasifica un vector en una u otra categoría resulta de gran interés para confiar en los resultados. Sería muy interesante implementar alguna técnica para visualizarlo.**

UEMC
universidad online



Muchas por vuestra atención