Fundamentos de la ciencia de datos

**Objetivo**

Esta actividad comprende una serie de ejercicios que han sido diseñados para reforzar la comprensión y aplicación de conceptos fundamentales en ciencia de datos y manejo de datos masivos. Cada ejercicio está orientado a revisar y profundizar en diferentes aspectos teóricos y prácticos de este campo:

* Comenzarás identificando y clasificando datos estructurados, semiestructurados y no estructurados, lo que te permitirá entender la variedad de formatos de datos y su adecuación a diferentes contextos analíticos.
* Posteriormente, explorarás un dataset específico para familiarizarte con la estructura de datos en formatos comunes como CSV, aprendiendo a interpretar la información contenida y su relevancia en análisis estadísticos y de salud pública.
* Investigarás y describirás datasets disponibles en plataformas de búsqueda, lo que te permitirá evaluar fuentes de información y entender su potencial de aplicación tanto en entornos académicos como comerciales.
* Analizarás en detalle un dataset desde una plataforma de datos abiertos, lo cual te ayudará a desarrollar habilidades para el manejo de datos científicos y su aplicación en investigaciones o desarrollos tecnológicos.
* Finalmente, examinarás cómo el big data se aplica en diferentes sectores industriales, permitiéndote visualizar la transversalidad y el impacto de la ciencia de datos en diversas áreas.
* Cada uno de estos ejercicios no solo te ayudará a afianzar tu conocimiento teórico, sino que también te proporcionará herramientas prácticas para abordar y resolver problemas complejos en la ciencia de datos, facilitando la toma de decisiones basada en datos en cualquier entorno profesional.

Recuerda, incluye en el informe a entregar al menos lo siguiente:

* Descripción de las fuentes de datos utilizadas.
* Discusión de los resultados obtenidos y argumentos que den valor a las conclusiones obtenidas.

**NOTA:** No está permitido el plagio entre compañeros de este o anteriores cursos, esta no es una actividad grupal. Tampoco está permitido el plagio de fuentes de Internet, libros, manuales o de cualquier otro tipo que no se citen apropiadamente. Entregar un ejercicio que incurra en los anteriores escenarios implicará automáticamente un 0 como calificación de la actividad, sin perjuicio de que se tomen medidas disciplinarias de acuerdo con la correspondiente normativa de la UNIR. Está permitido utilizar código, fracciones de código o ejemplos existentes en Internet u otras fuentes siempre que se **citen apropiadamente** y su uso se limite a partes minoritarias de la actividad.

Ejercicios por desarrollar

Ejercicio 1

Proporciona ejemplos de datos estructurados, semiestructurados y no estructurados para un escenario concreto, como, por ejemplo, una investigación relativa a una enfermedad en pacientes en función de la contaminación urbana.

Ejercicio 2

Vamos a explorar un dataset público descargable directamente en CSV.

Accede al portal de datos abiertos de la Junta de Castilla y León (España): <https://datosabiertos.jcyl.es/web/es/datos-abiertos-castilla-leon.html>

Busca el dataset «Situación epidemiológica coronavirus (COVID-19) en Castilla y León por provincias» que puedes encontrar en este enlace directamente:

Dirección General de Salud Pública - Consejería de Sanidad. (2020). *Situación epidemiológica coronavirus (COVID-19) en Castilla y León por provincias.* Junta de Castilla y León. <https://datosabiertos.jcyl.es/web/jcyl/set/es/salud/situacion-epidemiologica-coronavirus/1284940407131>

Descarga el dataset en formato CSV, ábrelo con MS Excel (por ejemplo) e indica el número de registros que contiene el dataset,cuántos campos por registro existen y qué significa cada campo.

Ejercicio 3

Accede al portal de Google de búsqueda dedatasets: <https://datasetsearch.research.google.com/>

Busca, por ejemplo, «covid» en el campo de búsqueda.

¿Cuántos datasets aparecen? Describe en términos cualitativos (sin entrar en detalles cuantitativos) los tres primeros datasets que aparezcan. ¿Qué entidades ofrecen cada uno de ellos? ¿Qué motivación tiene cada una de ellas para publicar dichos datos? ¿Qué utilidad se le podría dar a estos por parte de una empresa privada? ¿Y qué utilidad se le podría dar por parte de un centro de investigación?

Ejercicio 4

Accede al portal OpenML: <https://www.openml.org/>

Y, más concretamente, al siguiente dataset: <https://www.openml.org/d/42079>

¿Para qué puede ser empleado este dataset? ¿Cuántos registros incluye? ¿Cuántos campos por cada registro? ¿Qué significado tiene cada campo?

Ejercicio 5

¿A qué sector pertenece cada una de las siguientes aplicaciones de soluciones big data?

* A. Monitorización de las infecciones por hongos en cultivos.
* B. Detección de tumores mediante el análisis de imágenes con redes neuronales.
* C. Trazabilidad de viajeros en el transporte urbano.
* D. Pasarela de pago mediante criptomonedas.

Ejercicio 6

Busca tres ejemplos de implementaciones de sistemas big data que hayan permitido moverse a ciudades hacia el concepto de *smart cities*.

Ejercicio 7

Accede al Instituto Nacional de Estadística de España (INE) en su portal online y busca datos de evolución del PIB entre 2013 y 2017 (por ejemplo, con base en el SEC 2010, es decir, el Sistema Europeo de Cuentas de 2010). Descarga dichos datos y elabora una gráfica con Excel que represente la evolución del PIB a nivel nacional entre dichos años. ¿Qué gráfica utilizarías para ello?

Ejercicio 8

¿Cuál sería el resultado de convertir el siguiente archivo CSV a un archivo JSON con la misma información?

id;título;género;autor;año

0;"Lazarillo de Tormes";"Novela picaresca";"Anónimo";1554

1;"Don Quijote de la Mancha";"Novela de aventuras";"Miguel de Cervantes Saavedra";1605

2;"Ilíada";"Epopeya";"Homero";-500

3;"Cien años de soledad";"Realismo mágico";"Gabriel García Márquez";1967

Ejercicio 9

Identifica el tipo para cada uno de los siguientes datos:

1. Número de ovejas en un rebaño.
2. Registro de la temperatura de un sensor a lo largo del año.
3. Comentario en un artículo de un blog.
4. Estado *on/off* de un interruptor.
5. Peso de un púgil.
6. Dimensiones de un paquete.

Ejercicio 10

Plantea un ejemplo en el que describas cómo han de cumplirse las 7 V del big data.

**Extensión**

Extensión máxima de la actividad: un único informe de 10 páginas como máximo de extensión (Arial 11, interlineado 1,5). **El informe no debe incluir anexos, portada ni repetir el enunciado o la rúbrica de la actividad.** Cualquier página que incluyas será computada dentro de la extensión entregada. No se evaluará ninguna página a partir más allá de la 10 página. **No es posible adjuntar ningún archivo adicional.**