***Introducción a la Ciencia de Datos* – PEC1**

**Presentación**

Evaluación continua (EC) de los contenidos lectivos correspondientes:

* Bloque 1: Los datos.
* Contenido: *Fundamentos de ciencia de datos* (autor: Julià Minguillón)
* *La sociedad de la información.*
* *Datos, información, conocimiento, ¿sabiduría?*
* *¿Qué es un dato?*
* *Ciclo de vida de los datos.*
* Bloque 2: La ciencia de datos.
* Contenido: *Los roles, ámbitos y nombres de la ciencia de datos* (autor: Marçal Mora)
* *Origen y evolución de la ciencia de datos.*
* *El rol del científico de datos.*
* *Ámbitos de la ciencia de datos.*
* *Conceptos de ciencia de datos.*

**Objetivos y competencias**

* Construir una definición de ciencia de datos.
* Entender los conceptos habitualmente usados y relacionados en el contexto de la ciencia de datos.
* Entender las fases del ciclo de vida del dato.
* Entender la necesidad de poner en práctica otras competencias al trabajar en el papel de un científico de datos.

**Declaración de trabajo original (no plagio) de la o el estudiante**

Yo, *NombreEstudiante*, declaro que para realizar esta entrega... *(completad la frase con vuestras propias palabras)*

**Enunciado de la PEC**

**PEC1: ¿Ciencia en los datos?**

| ***Criterios de evaluación generales de la PEC.***   * *Aportación de alguna referencia externa que complemente o sustente los razonamientos que se expongan.* * *No omitir ninguno de los apartados de cada ejercicio.* * *Respetar la extensión de palabras señalada en cada enunciado.* * *Claridad en las respuestas y en los razonamientos.* * *Capacidad de síntesis.* * *Originalidad.* * *Sólo se permite el uso de herramientas de IA para*   *○ Búsqueda de información*  *○ Mejora de la ortografía, la gramática y la claridad de un texto.*  *○ Traducción de un texto. Cualquier otro uso no está permitido.*  ***Formato de entrega***   * *Las respuestas se entregarán en* ***formato PDF****.* * ***Las respuestas se subirán a la caja de Contenidos del aula correspondiente****, denominada “PEC1, Entrega estudiantes”.* * *El documento* ***no debe incluir el enunciado****, solo las respuestas.* * *El documento debe estar* ***estructurado*** *y el texto en un color que facilite la lectura (negro o azul oscuro).* * *La* ***fecha máxima*** *para entregar las respuestas es el* ***14 de abril de 2024, a las 23:59h****. No se corregirán las PECs que no cumplan este requisito, excepto en casos de fuerza mayor y debidamente justificados.* |
| --- |

**Ejercicio 1 (20% puntuación)**

**Enunciado**

En los dos primeros módulos de la asignatura, habéis visto qué es un **científico de datos** y sus principales responsabilidades. También habéis visto que existen diferentes roles, en función de los objetivos perseguidos: el *data* *scientist*, el *data engineer* y el *data analyst*.

Actualmente, la ciencia de datos sigue siendo un campo en constante evolución, por lo que han ido apareciendo **nuevos roles** que se adaptan a las necesidades específicas de las organizaciones y los avances tecnológicos, y a menudo en la práctica, los profesionales de la ciencia de datos realizan funciones que abarcan varios de estos roles.

Siguiendo el contenido del material “**Los roles, ámbitos y nombres de la**

**ciencia de datos**”, compartido en clase:

1. Responded:
   1. Teniendo en cuenta las tres disciplinas básicas que se combinan en el rol de científico o científica de datos, ¿qué dos disciplinas se combinan en el Aprendizaje Automático o Machine Learning?
   2. ¿Qué dos disciplinas se combinan en el desarrollo de software?
   3. ¿Por qué el perfil de Machine Learning a veces se considera ***peligroso***en una empresa?
   4. Explique brevemente uno de los diez puntos en los cuales se resume el trabajo diario en el rol de Data Scientist, **indicando además el número que ocupa en la lista que los recoge en los contenidos de referencia**.
   5. Según la propuesta que se hace en el documento del proceso de la ciencia de datos, ¿cuál sería el paso previo al Data Mining ? Explíquelo brevemente.
   6. ¿Cuál es el ámbito de aplicación de la Ciencia de datos de un algoritmo que puede aprender de múltiples perfiles y determinar que un individuo tiene el potencial de cometer un robo? ¿Puede llevar a la detención de una persona por un crimen «hipotético» al estilo de **The Minority Report**(Philip K. Dick, 1956)?
   7. ¿Cuál de los términos fundamentales de la Ciencia de datos se trata en el siguiente video: <https://www.youtube.com/watch?v=DQWI1kvmwRg>?¿Podrías explicarlo brevemente?
   8. Según el material compartido en el aula, ¿qué rol se define en clave de humor y cómo se define?

A continuación, vamos a ver un ejemplo práctico.

“En un proyecto de análisis de datos en salud, cada rol trabaja de manera independiente, demostrando la confianza en la capacidad de cada profesional para llevar a cabo sus tareas eficientemente. La Data Analyst inicia el proceso haciendo visualizaciones sin una revisión exhaustiva de la calidad de los datos, confiando en la eficacia de los sistemas de recopilación de datos de hospitales y laboratorios. El Data Engineer comienza a crear modelos sin realizar una limpieza previa de los datos recopilados de diversas fuentes, lo que le permite concentrarse directamente en la construcción de modelos predictivos para predecir diagnósticos médicos. La Data Scientist, a pesar de las posibles limitaciones en los datos provenientes de registros médicos y pruebas de laboratorio, lleva a cabo un análisis exploratorio sólido, considerando la situación como una oportunidad para identificar patrones complejos y desafiantes en la información de salud. Por último, el Machine Learning Engineer valora la oportunidad de mejorar la calidad de los datos y modelos al regresar al principio del proceso y preparar toda la infraestructura para el almacenamiento y el flujo de los datos, demostrando su capacidad para asegurar la adecuada preparación de los datos antes de su análisis.

Este enfoque, aunque aparentemente carece de coordinación y limpieza inicial de los datos, explota y resalta las habilidades y capacidades individuales de cada miembro del equipo en el análisis de datos en salud. Además, aumenta la velocidad del procesamiento de los datos y entrenamiento de los modelos ahorrando tiempo en pasos que no son importantes ni estrictamente necesarios.”

1. Analiza y comenta de manera crítica el ejemplo práctico expuesto teniendo en cuenta el material de los Bloques 1 y 2 de nuestro curso y los siguientes aspectos:
   1. El orden de los pasos y procesos descritos en el ejemplo.
   2. Las responsabilidades de cada rol.
   3. La importancia de los distintos pasos mencionados en el ejemplo.
   4. La interrelación entre los distintos roles así como “la aparentemente carencia de coordinación y limpieza inicial de los datos” que se señala en el ejemplo práctico.
2. ¿Cómo quedaría redactado el ejemplo si tuvieras que añadir algún cambio? Al margen de que quieras explicar o no los cambios, lo que se pide en esta pregunta es la versión del ejemplo práctico con los cambios que consideres necesarios basándote en su análisis crítico.

**Observación**: En la práctica, las funciones descritas anteriormente pueden ser asimiladas por diferentes roles, en función del tamaño de la empresa, del equipo y los roles que lo constituyen, por lo que vuestras respuestas deben considerar solamente el equipo que se ha planteado para el caso del ejemplo.

**Criterios de evaluación**

* El **apartado 1** se valorará con **1 punto** como máximo, **0.125 puntos** por cada rol bien definido.
* El **apartado 2** se valorará con **0.5 puntos** como máximo, **0.125** **puntos** por cada aspecto correctamente argumentado: orden lógico del proceso, responsabilidades de cada rol, importancia de los pasos, interrelación y cooperación entre los distintos roles.
* El **apartado 3** se valorará con **0.5 puntos** como máximo, **0.125** **puntos** por cada aspecto correctamente tratado en el ejemplo práctico: orden lógico del proceso, responsabilidades de cada rol, importancia de los pasos, interrelación y cooperación entre los distintos roles.
* La extensión máxima del apartado 1 es de **500 palabras**.
* La extensión máxima del apartado 2 es de **600 palabras**.
* La extensión máxima del apartado 3 es de **300 palabras**.
* Se valorará la capacidad de identificar los textos que justifican las respuestas.
* Todas las respuestas tienen que estar desarrolladas y argumentadas.

**Ejercicio 2 (30% puntuación)**

### **Enunciado**

En el primer módulo hemos visto las diferentes etapas del **ciclo de vida de los datos**, desde que se generan hasta que se publican. Por otra parte, en el segundo módulo, se ha presentado el **rol del científico de datos** y las tareas que realiza en su trabajo diario para dar respuestas a las preguntas planteadas.

La adopción de Inteligencia Artificial (IA) en medicina está revolucionando el diagnóstico rápido y la personalización del tratamiento. La cantidad de datos relacionados con la salud de las personas ha crecido de una manera muy rápida. Cada día es más difícil que el personal de salud pueda extraer información de relevancia usando métodos tradicionales de procesamiento de estos datos.

A diferencia de los métodos tradicionales, la IA analiza eficientemente grandes volúmenes de datos de salud, ofreciendo diagnósticos precisos y rápidos. En este contexto, la implementación de soluciones basadas en **IA y el uso adecuado del ciclo de vida de los datos** pueden abordar estos desafíos y mejorar significativamente el diagnóstico y tratamiento de enfermedades. En el siguiente enlace tenéis un caso de estudio a modo de ejemplo [consulta realizada el 09/03/2024]:

* [Inteligencia artificial al servicio de la salud del futuro](https://www.elsevier.es/es-revista-revista-medica-clinica-las-condes-202-articulo-inteligencia-artificial-al-servicio-salud-S0716864023000032)

A modo de resumen, algunos de los problemas específicos de los retos actuales de los sistemas de salud son:

* **Medicina Personalizada:** El desafío de personalizar tratamientos según la genética única y los factores ambientales de cada paciente, asegurando efectividad y minimizando riesgos.
* **Medicina Predictiva:** La dificultad de identificar con precisión a individuos en riesgo de desarrollar enfermedades antes de que estas se manifiesten, para permitir intervenciones preventivas tempranas.
* **Medicina Participativa:** Incrementar la implicación de los pacientes en su propio cuidado de salud, promoviendo una toma de decisiones informada y mejorando el cumplimiento del tratamiento.
* **Medicina Preventiva:** Fortalecer los esfuerzos de prevención de enfermedades a través de la detección temprana y la intervención precoz, enfrentando desafíos como el alcance y la aceptación de programas de cribado y vacunación.

En este contexto planteado, imaginad que un sistema de salud solicita sus servicios como científicos de datos con el objetivo de mejorar su atención médica. Con este fin, desarrollad los siguientes puntos:

1. Describe brevemente qué problema o problemas de los descritos anteriormente os gustaría solucionar. Establece claramente **una pregunta** **o una hipótesis de trabajo.**
2. **Captura de los datos.** Basándote en el caso de estudio proporcionado ¿Qué datos crudos utilizarías? ¿Dónde podrías obtenerlos? No hace falta que os limitéis exclusivamente a las fuentes de datos proporcionadas directamente por el sistema de salud.
3. **Almacenamiento y/o *Preprocessing*.** ¿Cómo y dónde los guardarías? ¿Les harías alguna modificación (cambio de formato, *data cleansing*, anonimización,etc.)?
4. **Análisis y/o visualización**. ¿Qué métricas, indicadores clave o datos calculados en base a datos crudos utilizarías para tu hipótesis? ¿Alguna idea de qué algoritmos u operaciones de las enumeradas en el caso de estudio proporcionado (a nivel muy general) usarías? (Por ejemplo: Una clasificación, regresión o agrupamiento (clustering))? ¿Cómo medirías el éxito del proyecto?
5. **Publicación**. Es importante que vuestros análisis generen conocimiento útil y/o valor en vuestro contexto de trabajo. ¿Qué harías con los resultados obtenidos? ¿Cómo los presentarías? ¿Quién sería el destinatario final de tus resultados?

**Criterios de evaluación**

* Cada **apartado** se valorará con **0.6 puntos** como máximo.
* Extensión máxima por apartado: **250 palabras**.
* Todas las respuestas tienen que estar **desarrolladas** y **argumentadas**.
* Las respuestas tienen que mostrar conexión con el temario teórico, de manera que se pueda percibir el aprendizaje desarrollado.
* **Siendo un ejercicio de asunción y simulación de un rol, los errores que pueda haber, en temas que no sean estrictamente vinculados con el temario teórico estudiado, no serán tenidos en cuenta**.

**Ejercicio 3 (30% puntuación)**

**Enunciado**

Como hemos visto a lo largo del segundo bloque de la asignatura, la **ciencia de datos** ha revolucionado, en diferentes ámbitos, la manera en cómo se aborda la **resolución de problemas** y la **toma de decisiones**. En este sentido, vamos a analizar dos casos de uso:

**Caso 1**: **Gestión de la energía en edificios**.

En la gestión de la energía en edificios, la ciencia de datos permite optimizar el uso energético y la participación en mercados de demanda flexible, como se muestra en el caso de "Flex2X" por Grid Edge. [consultas realizadas el 09/03/2024]:

* [Case Study: Artificial Intelligence for Building Energy Management Systems](https://www.iea.org/articles/case-study-artificial-intelligence-for-building-energy-management-systems)

**Caso 2**: **Agricultura de precisión.**

John Deere utiliza IA en la agricultura de precisión para aumentar el rendimiento de los cultivos mediante el **análisis de datos** y la automatización, demostrando cómo la tecnología puede llevar a una agricultura más eficiente y sostenible [consultas realizadas el 09/03/2024]:

* [Down to the Individual Grain: How John Deere Uses Industrial AI to Increase Crop Yields Through Precision Agriculture](https://www.databricks.com/blog/2021/07/09/down-to-the-individual-grain-how-john-deere-uses-industrial-ai-to-increase-crop-yields-through-precision-agriculture.html)

Considerando todo ello, **escoged uno de los dos casos presentados anteriormente y redactad un análisis** de entre 500 y 1000 palabras, con el apoyo de los ejemplos facilitados para cada caso o de otras fuentes, utilizando la **siguiente estructura**:

1. **Contexto** de la aplicación de la Ciencia de datos en el caso escogido.
2. **Objetivos y expectativas**. ¿Para qué usaron la Ciencia de Datos?
3. **Impacto** para la empresa y/o la sociedad.

Los enlaces suministrados son introductorios, de manera que podéis consultar más información *online* sobre el caso seleccionado, siempre y cuando citéis vuestras fuentes.

**Criterios de evaluación**

* Cada apartado se valorará con **1 punto** como máximo.
* Relacionar lo encontrado en el caso seleccionado con el material lectivo.
* Analizar las aplicaciones de Ciencia de datos.
* Contextualizar la ventaja competitiva que aporta la Ciencia de datos.
* La redacción que contenga **menos de 500 palabras o más de 1000 palabras se penalizará** con **0.2 puntos**.

**Ejercicio 4 (20% puntuación)**

**Enunciado**

Como habéis podido ver en el segundo módulo, los datos abiertos, u ***Open Data***, son una importante fuente de datos para investigadores de distintas áreas científicas, el periodismo de datos y para los Científicos de Datos.

Ahora, centrémonos en un recurso específico de datos abiertos [consulta realizada el 10-03-2024]:

* [Barcelona Open Data](https://opendata-ajuntament.barcelona.cat/es/node)

Analiza del portal y responded:

1. ¿Qué es Open Data Barcelona?

2. ¿Qué son los datos abiertos? ¿Cuáles son sus principales características?

Después de explorar y analizar el portal, responded:

1. ¿Cuántos datasets hay en el Catálogo de Open Data BCN? Describid el portal y los contenidos que ofrece, comentando al menos tres de las secciones del menú principal.



A continuación en la Clasificación por temas podremos encontrar el tema Territorio. A su vez, en este tema podremos encontrar datasets como el [Inventario de los separadores de carriles bici de la red de infraestructuras de la ciudad de Barcelona](https://opendata-ajuntament.barcelona.cat/data/es/dataset/infraestructures-inventari-separa-carrils-bici).

1. ¿Cuál es el grado de apertura del dataset? Comentad por qué motivo se le da ese grado de apertura(según la propia página), si hay en el dataset recursos con un formato asociado a datos jerárquicos y describid brevemente el esquema que se está utilizando, quién creó el esquema de clasificación, etc.



1. ¿Qué formatos de datos podemos encontrar en el portal? Comentad al menos tres formatos diferentes y su grado de apertura.

### **Criterios de evaluación**

* Los dos primeros apartados se valorarán con **0.25 puntos** cómo máximo respectivamente.
* El resto de apartados(del 3 al 5), se valorarán con **0.5 puntos** **cómo máximo** cada uno.
* Extensión máxima por pregunta: **350 palabras.**
* Explorar los portales de búsqueda de datos abiertos y comprender el contenido de estos.
* Comprender el significado de Open Data y su utilidad.
* Comprender el esquema de 5 estrellas.
* Comentar los distintos formatos de los datos.