PEC 1 (20% nota final)

Presentación

En esta Prueba de Evaluación Continuada se trabajan los conceptos generales del ciclo de vida de los datos, y se identifican y revisan sus características. También se trabajan los conceptos fundamentales del Web Scraping.

Competencias

En esta PEC se desarrollan las siguientes competencias del Máster de Data Science:

- Capacidad de analizar un problema en el nivel de abstracción adecuado a cada situación y aplicar las habilidades y conocimientos adquiridos para abordarlo y resolverlo.
- Capacidad para aplicar las técnicas específicas de tratamiento, almacenamiento y administración de datos

Objetivos

Los objetivos concretos de esta Prueba de Evaluación Continua son:

- Conocer el ciclo de vida de los datos y los principales tipos de datos.
- Aprender a aplicar los conocimientos adquiridos y su capacidad de resolución de problemas en entornos nuevos o poco conocidos dentro de contextos más amplios o multidisciplinarios.
- Desarrollar las habilidades de aprendizaje que permitan continuar estudiando de una manera que tendrá que ser en gran medida autodirigida o autónoma.
- Desarrollar la capacidad de busca, gestión y uso de información y recursos en el ámbito de la ciencia de datos.
- Entender la utilidad, la legalidad y algunas características del web scraping.

Descripción de la PEC a realizar

Ejercicio 1 [70%]

Después de leer el recurso “Calvo, M., Pérez, D., Subirats, L. (2019). Introducción al ciclo de vida de los datos.” contesta las siguientes preguntas con tus propias palabras:

1 ¿Qué perfil profesional dentro del ámbito de la ciencia de datos te llama más la atención? Explica su rol y cuáles serían los lenguajes de programación necesarios para ejercerlo (Máximo 100 palabras) [10%]

Líder de ciencia de datos (data science leader). Gestiona un equipo de analistas y científicos de datos. Como lenguajes de programación es necesario conocer R, SAS, Python, Matlab, Java.

2 ¿Qué técnica consiste en el conjunto de procesos que permiten corregir o eliminar aquellas muestras erróneas de una base de datos? Explica brevemente su definición y da un ejemplo (Máximo 100 palabras).[10%]

Limpieza. La limpieza de datos o data cleaning se considera uno de los pasos más importantes del preprocesado de los datos, ya que la calidad y la veracidad de los resultados van a depender en gran parte del correcto desarrollo de esta fase. Por ejemplo, limpiar los valores nulos del dataset.

3 A partir de los siguientes ejemplos, responde qué factor de calidad de datos está siendo afectado, da una breve definición para cada uno (Máximo 100 palabras por caso) [10%]:

-Un campo que indica que un alumno pasa o repite curso. Para pasar debe tener como máximo 1 asignatura suspendida. Un alumno con 2 asignaturas suspendidas y que pasa de curso sería un problema de qué tipo?

Consistencia. Es la ausencia de diferencia, al comparar dos o más representaciones de una cosa con su definición. La referencia es cada ítem, el ámbito es la base de datos y la unidad de medida es un porcentaje. Las dimensiones relacionadas son la exactitud, la validez y la unicidad.

-Un paciente tiene pautado un medicamento a las 8 AM, la enfermera de planta se lo entrega a las 8:15 y esta misma enfermera, tras la ronda de toda la planta, mecaniza la administración a las 9h.

Puntualidad. Se define la puntualidad (o atemporalidad) como el grado en que los datos representan la realidad desde un punto requerido en el tiempo. La referencia es el tiempo del evento real que ha estado obtenido y la medida es la diferencia en el tiempo.

-Porcentaje de direcciones incompletas/erróneas que utiliza una empresa de reparto de mensajería.
Completitud. Se define como la proporción de datos almacenados frente al potencial «100 % completo». La referencia son las reglas de negocio que definen lo que representa el «100 % completo»

-Un administrativo de una compañía teleoperadora registra la marca temporal de ciertas llamadas. Sin embargo, la marca temporal contiene únicamente la fecha y si la llamada se realizó en horario AM o PM.

Exactitud. Se define como el grado en que los datos describen correctamente el objeto o evento del «mundo real». Idealmente, la verdad del «mundo real» se establece a través de la investigación primaria.

4 En qué consiste el Análisis de datos perdidos, cuales son las técnica más usadas y da un ejemplo de dónde usarías algunas de estas técnicas (Máximo 200 palabras). [10%]

La denominación de datos perdidos o missing data se emplea cuando, para una variable u observación, no se tiene ningún dato. Este es uno de los problemas más comunes que se dan en el momento de la verificación de los datos, antes de realizar su limpieza. Pueden surgir por el mal funcionamiento de los dispositivos o procesos de captura, errores (olvidos) humanos o por errores de transmisión de datos.

Podemos destacar como técnicas:
Ignorar el atributo, Completado Manual, Completado Con Una Constante Global, Completado A Partir De Medidas Tendencia Central, Completado El Valor Más Probable.
Un ejemplo sería usar una regresión lineal para poder predecir a qué grupo pertenece cierto individuo utilizando el conjunto de variables de mi dataset y los datos aprendidos/clasificados con anterioridad.

5 Cuáles son los principales problemas causados por los valores extremos. Da un ejemplo de un valor extremo que pueda ser considerado falso, y un ejemplo que pueda considerarse real (Máximo 200 palabras). [10%]

1. Incrementando el error en la varianza reduciendo poder estadísticos, alterando así sus resultados.
2. Pueden afectar los resultados sobre correlaciones entre variables o regresiones
3. Sesgan los cálculos y estimaciones sobre un conjunto de datos al no pertenecer a la población de interés

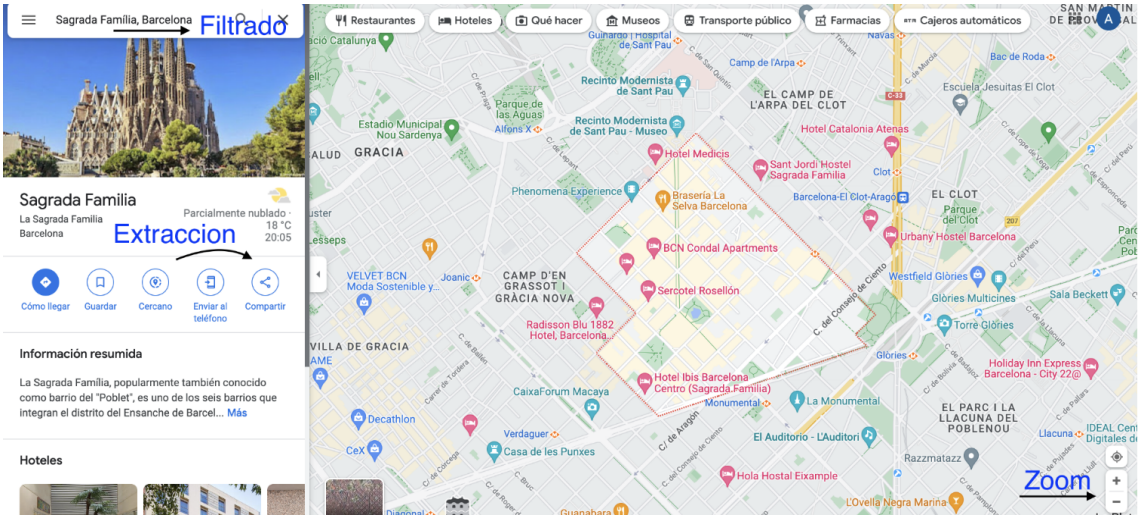
Como ejemplo tenemos un termostato que da como valor una temperatura de -300 grados centígrados. Y como ejemplo real tenemos la altura de una persona que llegue a los 2.20m, el cual si bien puede ser un valor outlier, podría ser un valor real y se debe estudiar su procedencia.

6 ¿Qué modelo o técnica estima una función o modelo a partir de una serie de datos de entrenamiento, con el objetivo de predecir posteriormente el resultado de nuevos datos desconocidos y cuál es su definición? (Máximo 100 palabras) [10%]

Modelo supervisados . Son métodos que se aplican cuando se disponen de datos en los que uno o varios atributos representan el objetivo del problema que se pretende resolver.
De esta manera, se diseñan modelos que buscan predecir los valores de nuevas variables de entrada correspondientes a dichos atributos a partir de las otras variables del conjunto de datos.

7 Menciona las siete tareas básicas que permiten un nivel más alto de abstracción para la visualización de datos, adjunta en imágenes ejemplos en donde se muestren al menos 3 de estas técnicas (Máximo 100 palabras).

Panorama general, acercamiento, filtrado, detalles a petición, relaciones, historial, extracción.



Ejercicio 2 [30%]

Después de leer el recurso “Subirats, L., Calvo, M. (2019). Web Scraping”, capítulos 1 y 6. Contesta las siguientes preguntas con tus propias palabras:

1 Indica un ejemplo de un sitio web que ofrezca un servicio de API, pero en el que igualmente resulte interesante aplicar web scraping, y justifica tu respuesta (Máximo 100 palabras). [15%]

https://www.idealista.com/
https://developers.idealista.com/access-request

Idealista es un sitio web que permite publicar inmuebles a la venta. La página, al ser tan popular, resulta interesante para el análisis de los precios en el sector inmobiliario por lo que puede ser muy interesante aplicar web scraping para obtener una perspectiva actualizada del sector. La página ofrece un servicio de API pero este debe ser aprobado por los administradores luego de evaluar el proyecto por lo que no es del todo conveniente.

2 ¿Qué son las trampas de araña? Indica tres ejemplos de algunas trampas que puedan encontrarse dentro de los sitios web. (Máximo 100 palabras). [15%]

Una trampa de araña es un conjunto de páginas web que se crean para provocar que un crawler se vea atrapado en un número infinito de peticiones. Ejemplos de trampa son los siguientes:

1. Páginas de calendario infinitas: en las que hay un calendario presente con enlaces a los meses anteriores y siguientes.
2. Enlaces defectuosos: enlaces que apuntan a URL defectuosas, generando aún más URL defectuosas.
3. URL con parámetros de consulta: Pueden dar lugar a infinitas URL únicas.

Recursos

Los siguientes recursos son de utilidad para la realización de la PEC:

Básicos

- Calvo, M., Pérez, D., Subirats, L. (2019). Introducción al ciclo de vida de los datos. Editorial UOC.
- Subirats, L., Calvo, M. (2019). Web Scraping. Editorial UOC.

Complementarios

- Lawson, R. (2015). Web Scraping with Python. Packt Publishing Ltd. Capítulo 1. Introduction to Web Scraping.
- Minguillón, J. (2016). Fundamentos de data Science. Editorial UOC.

Criterios de valoración

La ponderación de los ejercicios es la siguiente:

- Ejercicio 1.1: 10%
- Ejercicio 1.2: 10%
- Ejercicio 1.3: 10%
- Ejercicio 1.4: 10%
- Ejercicio 1.5: 10%
- Ejercicio 1.6: 10%
- Ejercicio 1.7: 10%
- Ejercicio 2.1: 15%
- Ejercicio 2.2: 15%

Se evaluará la precisión de los ejemplos así como lo respecto al número de palabras máximo establecido para cada pregunta. La idoneidad y claridad de las respuestas también será evaluada. No se pueden dar ejemplos que se mencionen directamente en los recursos proporcionados.

Formato y fecha de entrega

Hay que librar un único documento Word, Open Office o PDF (**preferiblemente este último**) con las respuestas a las preguntas.

Este documento se tiene que librar en el espacio de Entrega y Registro de AC del aula antes de las **23:59** del día **21 de marzo**. No se aceptarán entregas fuera de plazo.