22-10-2024

Captura y Preparación de Datos.

PEC1

Juan Luis Acebal Rico
GRADO DE CIENCIA DE DATOS APLICADA

Tabla de contenido

Parte teórica

Q1.

¿Qué representa la pirámide DICS? ¿Cuál es la diferencia entre datos, información y conocimiento?

Es un modelo que muestra la jerarquía de los datos, es decir, qué representa cada dato según el nivel donde esté en el modelo, siendo una jerarquía de valor e información del dato. Por ejemplo, en esta pregunta concreta, un **dato** sería 1985, **información** (quién, qué, dónde, y cuándo) sería el año de nacimiento con un contexto que sería la persona (con sus otros datos personales), y **conocimiento** (cómo) sería un conjunto de datos e información que den contexto y comprensión a la toma de decisiones, por ejemplo, las personas de 39 años (2024-1985) que hacen compras de suscripciones de plataformas de streaming. Por último, queda **sabiduría**(por qué), que, con mis palabras, es la capacidad de interpretar el conocimiento, con su contexto tomando decisiones, en base a la experiencia, formación, intuición, etc, es decir, la capacidad de extraer insights.

Q2.

¿Qué tipos distintos de datos podemos encontrar, según su estructura? ¿Y desde un punto de vista estadístico? Comenta muy brevemente cada tipo según estructura y a nivel estadístico y expón un ejemplo de cada tipo de dato.

Según su estructura podemos encontrar:

- Datos estructurados: Tienen una estructura clara: Un archivo CSV, una tabla de SQL
- No estructurados: No tienen una estructura clara: Un video, un email o un documento
- Semi estructurados: No tienen una estructura clara pero tienen etiquetas o marcadores propios de estos archivos que facilitan su análisis y su parseo: archivo JSON, XML.

Según un punto de vista estadístico:

- Datos cuantitativos: son datos numéricos, es decir, 160 cm. Además, podemos dividirlos en continuos o discretos.
    - Continuos: Altura: 160,33 cm, Saldo: 5643,23€
    - Discretos: Facturas: 1,2,3 (no hay facturas 1,5), tampoco hay 3,65 profesores colaboradores de una asignatura, etc.
- Datos cualitativos: también llamados categóricos, que representan un atributo, y se pueden dividir en nominales y ordinales
    - Nominales: Género=[‘Hombre’,’Mujer’], Estado_civil=[‘Casado/a’,’Soltero/a’, ’Viudo/a’,’Divorciado/a’]

- Ordinales: Son categorías que existen en orden lógico, por ejemplo, nivel de satisfacción. Podemos subdividirlos en:
  - Secuenciales: Edad
  - Divergentes: Votación a favor, en contra y abstención. Aquí sería el valor central abstención que es lo que caracteriza los datos cualitativos ordinales divergentes.
  - Cíclicos: Meses del año.

Q3.

Explica brevemente cuál es el ciclo de los datos y describe de forma breve las diferentes fases que lo conforman.

Es un proceso continuo que abarca desde la recolección hasta su análisis, almacenamiento o visualización:

- **Generación:** Es la fase de toma de datos, de distintas fuentes, y lugares, y con distintas tecnologías. Incluye dispositivos IoT, encuestas, datos de ventas, etc… Por ejemplo, el uso en una misma empresa de soluciones de datos que usan diferentes soluciones de datos.

- **Captura:** En esta fase los datos se recolectan (por ejemplo, con Airflow creando automatizaciones para su recolección)
- **Almacenamiento:** Se guardan los datos en bases de datos (por ejemplo, en un área stage con datos en bruto, o un DWH donde se pueden realizar análisis analíticos) o sistemas de almacenamiento.
- **Preprocesado:** Se procesa, se limpian datos, se normalizan, se eliminan valores duplicados, se convierten formatos, se eliminan fallas de tipo de dato, etc, incluso se pueden transformar datos que en un principio son no estructurados, tales como un documento, a datos semiestructurados o estructurados.
- **Análisis:** Se generan técnicas de estilo machine learning, o simplemente estadística para extraer información importante para encontrar o crear modelos predictivos o simplemente identificar tendencias relevantes.
- **Visualización:** Se generan gráficos para facilitar la comprensión. Por ejemplo un dashboard en Power BI o en Tableau.
- **Interpretación:** Se interpreta, se analiza el contexto y se puede entonces tomar decisiones, tales como la compra de inventario, o la creación de un departamento de riesgos, por ejemplo.

Q4.

¿Qué significa el concepto de "datos abiertos"? ¿Qué piensas al respecto? ¿Has encontrado alguna referencia a este a lo largo de tu vida profesional o de estudiante?

Los datos abiertos son datos que están disponibles al público, sin restricciones de uso, muchas veces bajo licencias de estilo CC0 o similar, y estos datos pueden ser utilizados, transformados, o redistribuidos libremente.

Mi opinión es que lo veo muy interesante, ya que, sino, nuestra profesión y el acceso al conocimiento estaría muy restringido, aunque, no siempre es posible encontrar los dataset que a uno le interesan, y veo en falta datos mas limpios, y más variedad. Si bien el acceso a datos sintéticos para abordar una idea o empezar un proyecto educativo o profesional es interesante, no es suficiente para ciertos contextos.

Por último, he utilizado y sobre todo, trabajado el concepto de datos abiertos el año pasado en la asignatura de introducción a la ciencia de datos, entre otras, además de tipología y fuentes de datos (si recuerdo bien), fueron datos del ayuntamiento de Barcelona, y también utilicé una página estatal que ofrecía datos abiertos y datos sintéticos también. Me parece muy interesante, pero lo veo insuficiente. Si recuerdo bien en introducción a ciencia de datos hemos utilizado para la predicción de la diabetes un dataset abierto.

También he utilizado datos de kaggle del uso de armas en EEUU programación en ciencia de datos y datos de elecciones en EEUU para la asignatura de bases de datos analíticas.

En mi vida personal he utilizado fuentes de datos abiertos del estilo datos demográficos, datos geográficos (concretamente para marcar divisiones territoriales en un mapa), y quizás eran abiertos también los datos de Yahoo Finance, al menos de uso permitido si, que usé para sacar datos de inflación vs SP500.

Laboralmente he usado datos geográficos (debido a que lo había usado como estudiante y colaboré en eso en concreto dada mi familiaridad), y ya. Los datos que usamos son el resto privados y confidenciales.

Q5.

¿Qué factores influyen en la calidad de los datos? Describe brevemente cada uno de ellos.

La calidad de los datos es una forma de expresar cómo los datos son exactos, completos, consistentes, puntuales, únicos, y válidos. Es decir, que valen para el propósito previsto, una calidad coherente con lo necesitado es importante para que las decisiones sean hechas y basadas en información precisa.

- **Exactitud:** Es como el nombre indica, que los datos sean precisos, deben ser correctos y representar la realidad que representan, y un ejemplo sería cuando, a todos nos ha pasado, en la dirección para algo, por ejemplo, de cliente, donde tiene que poner tu dirección, pone solamente tu localidad, o cuando en edad, pone 99 o 0. O sale la calle, pero no el número o el piso y letra.
- **Completitud:** Deben ser datos completos, una dirección sin el código postal y localidad no sirve.
- **Consistencia:** Los datos deben ser coherentes, si para una compañía telefónica soy Juan Luis Acebal Rico, y para otra soy Luis JA Rico, y pido una portabilidad, lo más seguro es que los sistemas lo rechacen ya que el titular no es quien lo solicitó.
- **Puntualidad:** Los datos deben estar actualizados, si yo he cambiado de dirección como cliente, tiene que existir un mecanismo para los clientes para actualizar la dirección o si no, la calidad de una base de datos de clientes no será buena.
- **Unicidad:** Es estar asegurado que yo no tendré 2 fichas de cliente, y que mis datos son únicos, o que un paciente no tiene su expediente médico asignado repetido y duplicado en 2 hospitales diferentes (dentro de la misma base de datos o sistema de bases de datos), es decir, si buscas Juan Luis Acebal Rico hay un paciente, y si buscas Juan L Acebal Rico es otro nombre, pero también otro expediente médico, pero ambos se refieren a mi persona, y ambos tienen datos que el otro no tiene, ya que aleatoriamente cada vez que voy al médico, el médico me busca, y actualiza datos médicos míos, en el expediente que entra antes.
- **Validez:** Se refiere a que el dato sea válido. No existe un DNI que tenga más de 1 letra, y esa letra tiene un algoritmo que la calcula, por ejemplo.