códigofacilito

Calidad de datos.

Dr. Niels Martínez Guevara

1 Datos y otros conceptos

2 introducción a la calidad de datos

3 Evaluación de calidad de datos

4 Mejora y manteniemiento de la calidad de los datos

¿Datos?

Fenómeno

Entendemos como fenómeno todo aquello que podemos percibir por nuestros sentidos y la razón [1]. Cuyas características podemos registrar en variables.

Datos estructurados

Son aquellos que se asocian con una variable por ejemplo velocidad, color, sabor, etc. Estas pueden ser cualitativas (descriptivas u ordinales) y cuantitativas [2].

Datos no estructurados

Son aquellos que no poseen una estructura generalizable, como los **textos**, las imágenes, videos, audios, etc [3].

[1] I. Kant (1978). «Crítica a la razón pura. Prólogo a la segunda edición». En: *Madrid. Alfaguara*

[2] J. D. Kelleher y B. Tierney (2018). *Data science*. MIT Press

[3] J. D. Kelleher y B. Tierney (2018). *Data science*. MIT Press

Introducción

Figura: Diagrama de Dikw enfocado en ciencia de datos [4]

[4] J. D. Kelleher y B. Tierney (2018). *Data science*. MIT Press

1 Datos y otros conceptos

2 introducción a la calidad de datos

3 Evaluación de calidad de datos

4 Mejora y manteniemiento de la calidad de los datos

El dato como un elemento de valor

- Un dato difícilmente va a poder generar información relevante sobre un tema determinado [5].
- Un conjunto de dato es el medio por el cual diferentes organismos pueden llegar buenas o malas decisiones [6].
- Es por eso que muchas instituciones ven a los datos como un valor activo, eficiente para la toma de decisiones (*Data Driven*).

[5] J. D. Kelleher y B. Tierney (2018). *Data science*. MIT Press

[6] L. Sebastian-Coleman (2022). *Meeting the challenges of data quality management*. Academic Press

¿Todo dato es bueno?

- Ciertamente los datos han cobrado mucha importancia en nuestra sociedad actual, sin embargo, esto ha traído consigo diferentes rumores o suposiciones que no son del todo ciertas en este ambito.
- Con el acercamiento de diferentes disciplinas a esta área emergente antes llevada en su totalidad por estos seres llamados estadísticos, podríamos experimentar ciertos errores ingenuos que podrían llevar a resultados no tan verídicos.
- Para ello existe el término *"garbaje in garbaje out"* el cual nos indica que si nuestros datos no son una aproximación certera del fenómeno de estudio debido a ciertos criterios que analizaremos más adelante, probablemente a las conclusiones que lleguemos no sean las más idóneas.

Malos datos, malos resultados

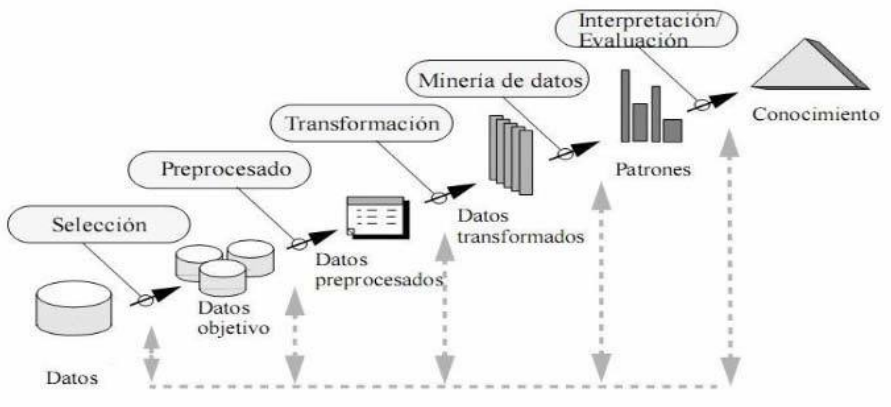

Figura: Modelo KDD, imagen recuperada [7]

7 J. Ramírez Pérez y R. Batista (abr. de 2015). «PROPUESTA DE RED CUBANA AURORA PARA LA COLABORACIÓN MÉDICA A TRAVÉS DE INFOMED UTILIZANDO UN ENFOQUE DE REDES SOCIALES». En

1. Datos y otros conceptos
2. introducción a la calidad de datos
3. Evaluación de calidad de datos
4. Mejora y manteniemiento de la calidad de los datos

KPI's y otros misceláneos

Un KPI (Key Performance Indicator, o Indicador Clave de Desempeño) es una métrica cuantificable que se utiliza para medir el rendimiento de los datos respecto a ciertos estándares de calidad. Estos indicadores ayudan a monitorear y evaluar la calidad de los datos a lo largo del tiempo, asegurando que cumplan con los requisitos necesarios para su uso efectivo en los procesos de negocio o análisis.

KPI's y otros misceláneos

- Ser Específico y Relevante: Debe estar alineado con los objetivos del negocio o la organización, midiendo un aspecto específico de la calidad de los datos que es crítico para el desempeño.
- Ser Cuantificable: Debe tener una métrica clara y numérica que permita el seguimiento y la comparación a lo largo del tiempo. Ser Alcanzable y Realista: Debe ser razonable y factible de alcanzar con los recursos disponibles.
- Estar Basado en el Tiempo: Debe tener un marco temporal definido para su medición, como mensual, trimestral, anual, etc.

Evaluacion de la calidad de datos

Las siguientes son las dimensiones de un proceso de calidad de datos:

- Precisión (Accuracy)Exactitud de los datos, es decir, qué tan cerca están de la realidad o la verdad. Datos precisos son aquellos que reflejan con precisión la información que intentan representar.
- Integridad (Completeness)Mide la totalidad de los datos. Un conjunto de datos completo es aquel que no tiene valores faltantes o huecos significativos. La integridad de los datos es crucial para obtener una visión completa y precisa.
- Validez (Validity)Indica si los datos están en conformidad con las reglas y estándares definidos. Los datos válidos cumplen con las restricciones y criterios establecidos para un conjunto de datos específico.

Evaluación de la calidad de datos

- Coherencia (Consistency)Se refiere a la uniformidad de los datos a lo largo del tiempo y entre diferentes conjuntos de datos. Datos coherentes no presentan contradicciones o discrepancias cuando se comparan entre sí.
- Unicidad (Uniqueness)Evalúa si no hay duplicados en los datos. Los datos únicos garantizan que cada entidad o elemento esté representado solo una vez en un conjunto de datos.
- Oportunidad (Timeliness)Se refiere a la actualidad de los datos. La información oportuna es aquella que está disponible cuando se necesita, sin demoras innecesarias.
- Aptitud (Fitness)Este aspecto evalúa la relevancia y utilidad de los datos para el propósito previsto. Los datos deben ser adecuados y aplicables a los objetivos específicos de la organización o del análisis que se esté llevando a cabo.

Evaluación de la calidad de datos

Los siguientes son algunas de las implementaciones para poder realizar validaciones de la calidad de los datos:

- Great Expectations: Great Expectations es una biblioteca open-source para la validación de datos. Permite definir, documentar y validar expectativas sobre los datos, garantizando la calidad y consistencia en proyectos de ciencia de datos y análisis.
- Pandera: Pandera es una biblioteca de validación de datos para estructuras de datos en Python, especialmente diseñada para trabajar con DataFrames de pandas. Permite definir esquemas y reglas de validación para asegurar la conformidad de los datos.
- Dora: Dora es una libreria python diseñada para automatizar exploración de datos y realizar analisis de datos exploratorios

Evaluación de datos

Debemos recordar que estas librerías al ser openSource en algunas ocasiones dependen de la comunidad para su actualización y mantenimiento, como es el caso de Pandas Profiling la cual ha pesar de ser muy util, se han modificado las librerías que utilizan dificultando su uso. Aunque siempre podemos analizar la naturaleza de nuestros datos a partir de diferentes scripts, funciones y visualizaciones.

1 Datos y otros conceptos

2 introducción a la calidad de datos

3 Evaluación de calidad de datos

4 Mejora y manteniemiento de la calidad de los datos

Estrategias para mejorar la calidad de los datos

- Implementar estrategias, políticas y procedimientos adecuados para la entrada de datos, si podemos tener unos datos adecuados en primer lugar esto ahorraría mucho trabajo en el curado de los datos, establecer restricciones en los sistemas de captura o establecer protocolos de llenado mejorara significativamente los datos.
- Definir procesos de limpieza y normalización de los datos, sí conocemos el contexto de nuestro fenómeno de estudio, esto facilita el poder implementar scripts para los procesos de curado de los datos.
- Automatización de procesos enfocados a la detección de la calidad de los datos. Si conocemos el contexto del problema podemos evitar que se registren errores o avisar al usuario.

Mantenimiento continuo de la calidad

- Definición de controles y auditoria sobre los datos de forma periódica.
- Capacitación del personal y definición de una cultura de calidad de los datos dentro de la organización.
- *"Los datos como el agua tienen memoria..."*

Bibliografía I

Kant, I. (1978). «Crítica a la razón pura. Prólogo a la segunda edición». En: *Madrid. Alfaguara*.

Kelleher, J. D. y B. Tierney (2018). *Data science*. MIT Press.

Ramírez Pérez, J. y R. Batista (abr. de 2015). «PROPUESTA DE RED CUBANA AURORA PARA LA COLABORACIÓN MÉDICA A TRAVÉS DE INFOMED UTILIZANDO UN ENFOQUE DE REDES SOCIALES». En.

Sebastian-Coleman, L. (2022). *Meeting the challenges of data quality management*. Academic Press.