Calidad de Datos e Información

Laboratorio 2023

Esta tarea se divide en 2 partes y, para el desarrollo de cada una de ellas, se consideran 2 datasets (desde ahora DS1 y DS2), los cuales serán descriptos en un documento aparte. A continuación se presenta la descripción de cada una de las partes del laboratorio.

PARTE 1

En esta parte de la tarea se pide realizar todas las tareas de gestión de calidad de datos vistas en clase, esto implica:

1. Realización de las tareas de Data Profiling (usando las herramientas *DataCleaner* o un notebook de *Python* con la biblioteca *Pandas*). Los estudiantes determinarán qué tan exhaustivas deben ser estas tareas.
2. Especificación del Modelo de Calidad de Datos.
3. Diseño de la base de Metadatos de Calidad.
4. Ejecución de la medición
5. Análisis de los resultados

NOTA: En esta parte de la tarea los grupos impares trabajaran con el DS1 y los grupos pares trabajaran con el DS2.

PARTE 2:

En esta parte de la tarea se pide aplicar la metodología de gestión de calidad de datos vista en clase (CaDQM). En este caso, nos concentraremos en la ejecución de las 6 primeras etapas, las cuales implican:

1. Definición del contexto de los datos.
2. Análisis de los datos.
3. Análisis de los requerimientos de usuarios.
4. Especificación del Modelo de Calidad de Datos
5. Diseño de la base de Metadatos de Calidad, ejecución de las métricas de medición y almacenamiento de los resultados obtenidos.
6. Evaluación de la calidad de los datos.

NOTA: En esta parte de la tarea los grupos impares trabajaran con el DS2 y los grupos pares trabajaran con el DS1.

INFORME

El informe final del laboratorio debe contener, al menos la siguiente información:

- **Introducción**
  Breve introducción del trabajo realizado y organización del documento.
- **PARTE 1**
  - **Introducción**
    Breve descripción de la forma de trabajo, como justificación de decisiones tomadas.
  - **Descripción general de la fuente de datos**
    Descripción general de las fuentes de datos (realidad a la que está asociada, archivos, formatos de datos, etc.).
  - **Data Profiling**
    Se debe entregar el análisis general de las fuentes, examinando la estructura, relaciones y volumen de datos. Se pretende tener una visión global de las fuentes de datos, evaluando estadísticas interesantes que puedan ser útiles para definir las dimensiones y factores de calidad a medir y los datos prioritarios.
  - **Especificación del Modelo de Calidad de Datos**
    El modelo de calidad de datos debe contener la siguiente información: dimensiones, factores, métricas y métodos de medición. Finalmente, deberán implementarse todas las métricas. Para acotar el tamaño del modelo de calidad, se deben plantear, como máximo, 6 factores para cada dimensión de calidad y 3 métricas para cada factor de calidad propuesto (proponga las más relevantes según su criterio). Además, se sugiere incorporar al modelo algunas agregaciones de medidas que consideren útiles para el análisis de la calidad.

    Por otro lado, si en algún caso lo considera relevante, puede incluir un análisis y/o justificación de las dimensiones y factores de calidad elegidos, así como de los datos que se consideraron prioritarios, comentando porqué son apropiados e interesantes para la realidad dada.

Para las métricas y métodos se pueden utilizar diccionarios como referencia para errores de digitación, y referenciales que se considerarán como la realidad, para las entidades que sea posible.

- **Diseño de la base de Metadatos de Calidad**

En esta etapa se diseña el esquema de Metadatos de Calidad a utilizar para registrar las medidas de calidad.

- **Análisis de resultados**

Se debe presentar un análisis de los resultados obtenidos en la evaluación de la calidad de los datos.

- **PARTE 2**

  - **Introducción**

Breve descripción de la forma de trabajo, como justificación de decisiones tomadas.

  - **Descripción general de la fuente de datos**

Descripción general de las fuentes de datos (realidad a la que está asociada, archivos, formatos de datos, etc.).

  - **Ejecución de la Metodología de Gestión de Calidad de Datos dependiente del Contexto**

Para las 6 primeras etapas de la metodología, se deben documentar todas las entradas, justificando todas las decisiones tomadas, y las salidas de cada etapa. En particular, se deben seguir los formatos de definición del contexto y de especificación del modelo de calidad vistos en clase. También las salidas de cada etapa de la metodología deben ser presentadas en el formato visto en clase.

- **Conclusiones**

En esta sección se deben documentar las conclusiones: i) de cada una de las partes, ii) generales, y iii) personales sobre el trabajo, análisis realizados.

- **Anexo**

Como anexo del informe, se deberá entregar las respuestas a un cuestionario sobre la metodología de gestión de calidad (CaDQM). Dicho cuestionario, será publicado a la brevedad.