INFORME

CIENCIA DE DATOS

CURSO: INTRODUCCIÓN ALA CIENCIA DE DATOS

AUTOR: YERALDINNE RUIZ COGOLLO

BRAYAN RAMOS

MARÍA DILSO

DIEGO SALAS

EMILIO PLAZA

DOCENTE: ALEXANDER TOSCANO

UNIVERSIDAD DE CÓRDOBA

LIC. INFORMÁTICA

MONTERÍA

2024

CIENCIA DE DATOS

LIMPIEZA DE DATOS

La limpieza de datos, también conocida como depuración de datos, es un proceso crucial dentro del ciclo de vida de la gestión y análisis de datos. Su objetivo principal es asegurar que los datos sean precisos, consistentes y utilizables, eliminando errores, inconsistencias o datos irrelevantes que puedan afectar los resultados de los análisis. Este proceso es especialmente importante en entornos donde los datos provienen de diversas fuentes, lo que puede aumentar la probabilidad de errores o incoherencias.

IMPORTANCIA

La limpieza de datos es fundamental para asegurar la calidad y fiabilidad en cualquier análisis o proceso de toma de decisiones basado en datos. Uno de sus principales beneficios es que mejora la calidad de los datos al eliminar errores, duplicados, valores atípicos e inconsistencias en el formato, lo que asegura que la información sea precisa y esté completa. Además, aumenta la precisión de los análisis, ya que trabajar con datos sucios o incompletos puede llevar a conclusiones erróneas y a decisiones inexactas. Un conjunto de datos limpio garantiza que los resultados del análisis representen de manera más fiel la realidad y sean aplicables a la resolución de problemas o a la planificación estratégica.

Asimismo, la limpieza de datos reduce sesgos y errores. Los datos sin depurar pueden contener valores que distorsionen los resultados, como registros mal etiquetados o información incorrecta. Este proceso ayuda a identificar y corregir estos problemas, reduciendo el riesgo de llegar a conclusiones equivocadas. También permite una mejor integración de los datos, especialmente si provienen de múltiples fuentes, lo que facilita su estandarización y compatibilidad para análisis más avanzados.

PROBLEMAS COMUNES

-Datos faltantes: Este problema puede surgir debido a errores en la recolección de datos, omisiones por parte de los encuestados o fallos en los sistemas de captura. La presencia de datos faltantes puede llevar a sesgos en el análisis y reducir la representatividad de los resultados, afectando así la calidad de las conclusiones que se puedan extraer.

- Duplicados: Este problema suele surgir al combinar datos de diferentes fuentes sin un proceso adecuado de deduplicación.

-Errores tipográficos y de formato: Estos problemas pueden ser causados por la introducción manual de datos o errores durante el proceso de migración de datos. La existencia de estos errores puede resultar en problemas de categorización y en la incapacidad de unir o comparar datos correctamente, afectando así la calidad del análisis.

-Valores atípicos (outliers):Este problema puede ser causado por errores en la medición, entrada de datos incorrecta o, en algunos casos, por variaciones naturales en los datos. La presencia de valores

atípicos puede afectar los resultados del análisis estadístico, llevando a conclusiones incorrectas o engañosas.

HERRAMIENTAS

Microsoft Excel es una herramienta ampliamente conocida y utilizada en la gestión de datos. Aunque no está diseñada exclusivamente para la limpieza de datos, ofrece funciones muy útiles para esta tarea, especialmente en conjuntos de datos pequeños o medianos. Excel permite la eliminación de duplicados, el filtrado y la ordenación de datos, la validación para asegurarse de que los valores sigan un formato o rango específico, y la utilización de funciones de búsqueda y reemplazo para corregir errores. Además, sus tablas dinámicas son útiles para detectar patrones y anomalías en los datos.

OpenRefine, antes conocido como Google Refine, es una herramienta de código abierto enfocada en la limpieza y transformación de grandes volúmenes de datos desordenados. Está diseñada para estandarizar datos de diferentes fuentes, corregir errores y transformar datos no estructurados como JSON o XML. Su capacidad para detectar y corregir inconsistencias de manera eficiente la convierte en una excelente opción para proyectos que requieren una limpieza exhaustiva y manipulación de datos complejos.

Trifacta es una plataforma basada en la nube diseñada específicamente para la limpieza, transformación y preparación de datos. Ofrece una interfaz visual intuitiva que permite a los usuarios detectar automáticamente problemas en los datos, como valores faltantes o duplicados. Trifacta facilita la preparación de datos para análisis posteriores, integrándose con varias fuentes de datos y permitiendo la creación de flujos de trabajo automatizados para grandes volúmenes de información.

EJEMPLO

Un ejemplo típico de la aplicación de limpieza de datos se da en una campaña de marketing por correo electrónico. Supongamos que una empresa ha recopilado datos de clientes a través de varias fuentes, como formularios de inscripción, encuestas y bases de datos internas. Antes de lanzar la campaña, es crucial limpiar los datos para garantizar que la información sea precisa y que los correos lleguen de manera efectiva a los destinatarios correctos.

Proceso de limpieza de datos

Eliminación de duplicados: Es común que un mismo cliente se registre en varias ocasiones en diferentes plataformas. Para evitar enviar correos duplicados, se identifican y eliminan los registros repetidos.

Corrección de errores tipográficos: En los datos de contacto, pueden existir errores en los correos electrónicos (por ejemplo, "john.doe@gamil.com" en lugar de "john.doe@gmail.com"). Se corrigen estos errores mediante técnicas de búsqueda y reemplazo.

Normalización de formatos: Los nombres de los clientes pueden aparecer en diferentes estilos (por ejemplo, "Juan Pérez" o "juan pérez"). Se estandarizan los formatos para asegurar uniformidad.

Imputación de datos faltantes: Si hay campos incompletos (como direcciones de correo faltantes), se decide cómo manejarlos, ya sea eliminando los registros incompletos o intentando recuperar la información a través de otros medios..