GOBIERNO DE
EL SALVADOR

SECRETARÍA DE
INNOVACIÓN

CIENCIA DE DATOS

Agenda
Sesión 7/18

Limpieza y procesamiento de datos

-Definición

- Librerías de Python

- Requisitos que debe cumplir un dato

- Proceso de Limpieza de Datos

- Métodos más usados

- Repaso

    - Tweepy

    - MongoDB Atlas

    - Migración de Datos en MongoDB

"La ciencia más útil es aquella cuyo fruto es el más comunicable"

Leonardo Da Vinci

Limpieza y procesamiento de datos

Es el primer paso mas critico en cualquier proyecto de inteligencia artificial y el Machine Learning.

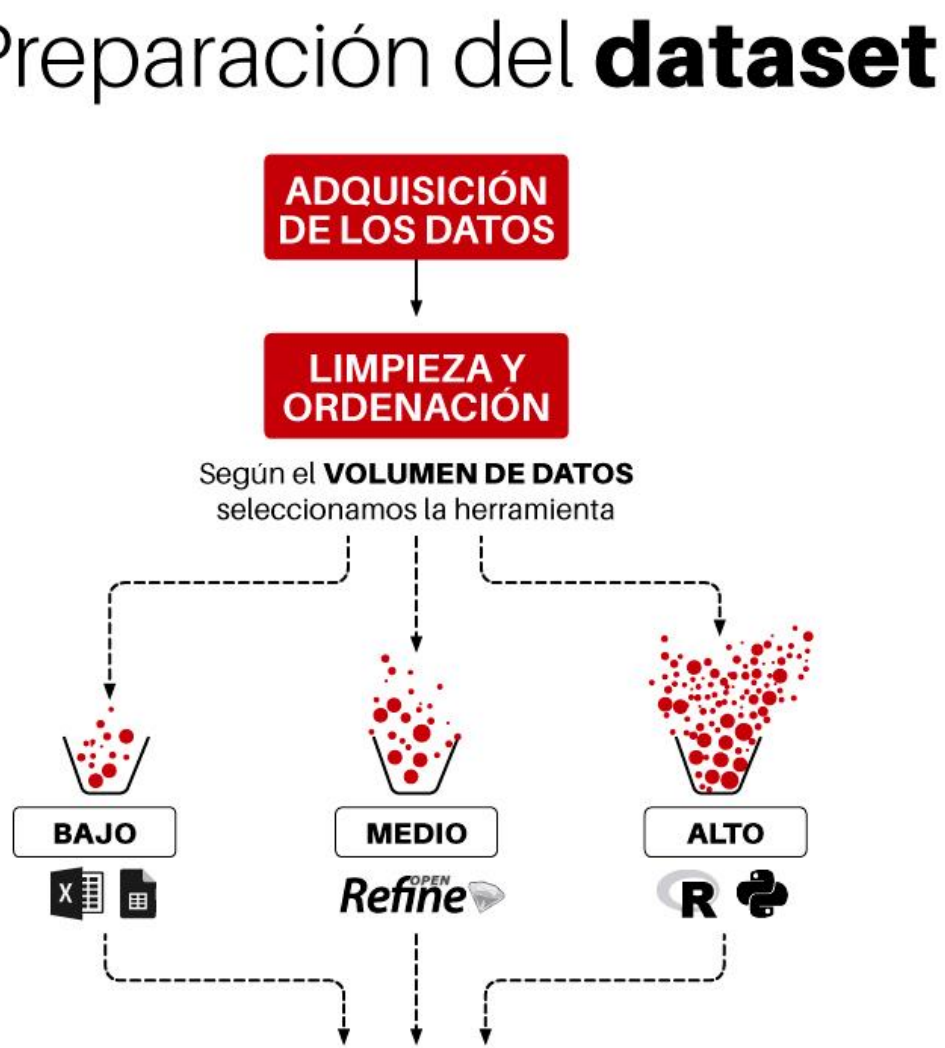
Es de los procesos mas importantes en el momento de hacer análisis de datos.

Limpieza y procesamiento de datos

Es el proceso de detectar, corregir, eliminar registros corruptos o inexactos de un conjunto de registros, tablas o bases de datos.

Se refiere también a la identificación de partes incompletas, incorrectas, inexactas o irrelevantes de los datos para su posterior sustitución, modificación, eliminación de los datos sucios o poco precisos

Importación de datos



Limpieza y procesamiento de datos

Librerías de Python para el preprocesamiento de los datos

Numpy

Librería generar operaciones matemática

Pandas

Importante para importar y gestionar conjuntos de datos

Matplotlib

Nos ayuda en la parte visual para desarrollo de graficas

Limpieza y procesamiento de datos

Limpieza y Procesamiento de datos en Python mas adelante

1. Importación de datos
2. Analizar los datos
3. Preprocesar los datos
4. Separar datos para entrenamiento y prueba

Limpieza y procesamiento de datos

La limpieza de datos (en inglés data cleansing o data scrubbing) es el acto de descubrimiento y corrección o eliminación de registros de datos erróneos de una tabla o base de datos. El proceso de limpieza de datos permite identificar datos incompletos, incorrectos, inexactos, no pertinentes, etc. y luego substituir, modificar o eliminar estos datos sucios ("data duty"). Después de la limpieza, la base de datos podrá ser compatible con otras bases de datos similares en el sistema.

Requisitos que debe cumplir un dato

La calidad de datos debe cumplir con los siguientes requisitos:

Exactitud:

Los datos deben cumplir los requisitos de integridad, consistencia y densidad.

Integridad:

Los datos deben cumplir los requisitos de Entereza y validez.

Entereza:

Alcanzado por la corrección de datos que contienen anomalías.

Requisitos que debe cumplir un dato

Validez:

Alcanzado por la cantidad de datos que satisfacen las restricciones de integridad.

Consistencia:

Alcanzado por la corrección de contradicciones y anomalías sintácticas.

Uniformidad:

Relacionado con irregularidades.

Requisitos que debe cumplir un dato

Densidad:

Conocer el cociente de valores omitidos sobre el número de valores totales.

Unicidad:

Relacionado con datos duplicados.

Proceso de Limpieza de Datos

Proceso de Limpieza de Datos

Auditoría de Datos:

Los datos son revisados con el empleo de métodos estadísticos de descubrir anomalías y contradicciones. Esto tarde o temprano da una indicación de las características de las anomalías y sus posiciones.

Proceso de Limpieza de Datos

Definición de Workflow (Flujo de Trabajo): La detección y el retiro de anomalías son realizados por una secuencia de operaciones sobre los datos sabidos como el workflow. Para alcanzar un workflow apropiado, se debe identificar las causas de las anomalías y errores. Si por ejemplo encontramos que una anomalía es un resultado de errores de máquina en etapas de entrada de datos, la disposición del teclado puede ayudar en la solución de posibles problemas.

Proceso de Limpieza de Datos

Ejecución de Workflow: En esta etapa, el workflow es ejecutado después de que su especificación es completa y su corrección es verificada. La implementación del workflow debería ser eficiente aún sobre los juegos grandes de los datos que inevitablemente plantean una compensación, porque la ejecución de la operación limpiadora puede ser cara.

Proceso de Limpieza de Datos

Post-Proceso y Control: Los datos que no podían ser corregidos durante la ejecución del workflow deberán ser corregidos manualmente, de ser posible. El resultado es un nuevo ciclo en el proceso de limpieza de datos donde los datos son revisados nuevamente para ajustarse a las especificaciones de un workflow adicional y realizar un tratamiento automático.

Métodos más usados

Métodos más usados

Análisis

El análisis en la limpieza de datos, es realizado para la detección de errores de sintaxis. Un analizador gramatical decide si una cuerda de datos es aceptable dentro de la especificación de datos permitida. Esto es similar al modo que un analizador gramatical trabaja con gramáticas y lenguas.

Transformación de Datos:

La Transformación de Datos permite al trazar un mapa de datos, en el formato esperado. Esto incluye conversiones de valor o funciones de traducción así como normalización de valores numéricos para conformarse a valores mínimos y máximos.

Métodos más usados

Eliminación de duplicados:

La detección de duplicados requiere un algoritmo para determinar si los datos contienen representaciones dobles de la misma entidad. Por lo general, los datos son ordenados por un dato "llave" o "pivote" que permite la identificación más rápida.

Métodos más usados

Método Estadístico:

Incluye analizar los datos usando promedios, desviación estándar, rangos, o algoritmos de cluster, este análisis se realiza por expertos que identifican errores. Aunque la corrección de datos sea difícil ya que no saben el valor verdadero, pueden ser resueltos poniendo los valores a un promedio u otro valor estadístico. Los métodos estadísticos también pueden ser usados para manejar los valores que fallan, que pueden ser substituidos por uno o varios valores posibles que por lo general son obtenidos por algoritmos de aumento de datos extensos.

Desafíos y Problemas

Desafíos y Problemas

Corrección de Error y pérdida de información:

El mayor desafío dentro de la limpieza de datos es la corrección de valores, pues incluye el quitar duplicados y entradas inválidas. En muchos casos, la información disponible sobre tales anomalías es limitada e insuficiente de determinar las transformaciones necesarias o correcciones abandonando la tachadura de tales entradas como la única solución. La eliminación de datos aunque, conduce a la pérdida de información que puede ser en particular costosa si hay una cantidad grande de datos suprimidos.

Desafíos y Problemas

Mantenimiento de Datos Limpiados:

La limpieza de datos es cara y el tiempo consumido es grande. Después de haber realizado la limpieza de datos y el alcanzar una colección de datos sin errores, uno querría evitar la relimpieza de datos íntegramente después de que se realizan algunos cambios en la base de datos. El proceso sólo debería ser repetido sobre los valores que se han cambiado, esto significa, que debemos guardar un linaje limpiador que requiere una eficiente colección de datos y técnicas de administración de datos.

Desafíos y Problemas

Limpieza de Datos en Entornos virtualmente Integrados:

En Fuentes prácticamente integradas como DiscoveryLink de la IBM, la limpieza de datos tiene que ser realizada siempre con acceso de datos de diferentes fuentes, con una considerable disminución el tiempo de respuesta y la eficacia.

Desafíos y Problemas

Limpieza de datos en el Framework:

En muchos casos no será posible llegar a un completo mapa de limpieza de datos, que guíe el proceso por adelantado. Esto hace que la limpieza de datos sea un proceso iterativo que implica la exploración significativa y la interacción que puede requerir un framework, es decir, un marco que incluya una colección de métodos para la detección de errores y la eliminación además de la revisión de datos. Esto puede ser integrado con otras etapas informáticas como la integración y el mantenimiento

Repaso General

GOBIERNO DE
EL SALVADOR

SECRETARÍA DE
INNOVACIÓN