TIC

LECCIÓN 1

Lección 2: Limpieza de datos

La limpieza de datos implica identificar, corregir o eliminar errores, inexactitudes o inconsistencias en conjuntos de datos. Es esencial para garantizar la calidad y la integridad de los datos antes de realizar análisis o modelos predictivos. Algunos pasos comunes en el proceso de limpieza de datos son: Identificar y manejar valores faltantes: Los valores faltantes son comunes en conjuntos de datos. Pueden ser eliminados, imputados (es decir, estimados o remplazados) con técnicas como la media, la mediana o modelos predictivos.

Manejar valores atípicos: Los valores atípicos pueden distorsionar el análisis. Pueden ser identificados mediante técnicas estadísticas y luego tratados mediante eliminación, transformación o imputación.

Corregir errores de datos: Esto implica identificar y corregir errores tipográficos, errores de formato o inconsistencias en los datos.

Estandarización y normalización de datos: Esto implica convertir los datos en un formato uniforme para facilitar el análisis. Por ejemplo, convertir todas las fechas al mismo formato o estandarizar unidades de medida.

Detección y manejo de duplicados: Identificar y eliminar registros duplicados en los conjuntos de datos.

Manejo de datos inconsistentes: Esto puede incluir la reconciliación de discrepancias entre diferentes fuentes de datos o la corrección de inconsistencias lógicas en los datos.

Validación de datos: Verificar la precisión y la coherencia de los datos en función de reglas predefinidas o conocimiento experto.

Normalización de variables categóricas: Convertir variables categóricas en un formato adecuado para su análisis, como la codificación one-hot encoding.

Reducción de dimensiones: En conjuntos de datos de alta dimensionalidad, reducir la cantidad de variables puede ayudar a mejorar la calidad de los datos y a acelerar los algoritmos de análisis.

Documentación del proceso de limpieza: Es importante documentar todas las acciones realizadas durante el proceso de limpieza de datos para garantizar la reproducibilidad y la transparencia del análisis.

En los procesos de limpieza y de reducción de dimensiones a veces es necesario la eliminación de columnas y filas ya que representan valores que no son convenientes tratar. Como es el caso de datos con mal formato, registros que están mal formados, outliers sin sentido o incluso, en ocasiones, datos que tienen una frecuencia inusualmente alta al compararlos con otros valores. Siempre es necesario analizar el por qué sucede algo y no dar por sentado que la distribución tiene un comportamiento inusual.