Introducción al Big Data

Sesión 18 -Limpieza de los datos

José Fuentes

25 September 2023

Introducción

La limpieza de datos es un proceso fundamental en el campo de la ciencia de datos.

- A medida que la disponibilidad de datos se ha vuelto cada vez más abundante, se ha reconocido la necesidad de garantizar la calidad y confiabilidad de los conjuntos de datos utilizados en análisis y modelado.
- La limpieza de datos se refiere al conjunto de técnicas y procedimientos destinados a identificar, corregir y eliminar errores, inconsistencias y valores atípicos en los datos.
- El objetivo principal de la limpieza de datos es mejorar la calidad y la integridad de los datos, asegurando que sean precisos, completos y consistentes.

Fases de la Limpieza de Datos

- **Identificación de problemas:** se analizan los datos en busca de posibles errores, inconsistencias o valores atípicos.
- **Evaluación de la calidad de los datos:** se determina la calidad de los datos en términos de precisión, completitud y consistencia.
- **Tratamiento de valores faltantes:** se decide cómo manejar los valores faltantes en los datos, ya sea eliminándolos, imputándolos o descartando las observaciones correspondientes.

Fases de la Limpieza de Datos...

- Eliminación de valores atípicos: se identifican y eliminan los valores atípicos que pueden afectar negativamente los análisis o modelos.
- Corrección de errores de formato: se corrigen los errores de formato, como fechas mal ingresadas o nombres mal escritos.
- Verificación de la coherencia y la integridad: se verifica la coherencia y la integridad de los datos para garantizar que sean lógicamente consistentes y libres de duplicados.

Importancia de la Limpieza de Datos

- El éxito de la limpieza de datos se puede medir en términos de la calidad mejorada de los datos, la confiabilidad de los resultados, la precisión de los análisis y la eficiencia en el uso de los datos.
- La limpieza de datos desempeña un papel crucial en la obtención de información precisa y confiable a partir de los datos, allanando el camino para una toma de decisiones informada y basada en evidencia.

Identificación de Problemas

En esta fase, se examinan los datos para identificar problemas potenciales, como valores faltantes, errores de formato, valores atípicos o inconsistentes.

- Supongamos que estás trabajando con un conjunto de datos que contiene información sobre transacciones financieras.
- Durante la fase de identificación de problemas, observas que hay una columna llamada "Monto" que debería contener los valores de las transacciones realizadas.
- Sin embargo, al examinar más de cerca los datos, descubres algunos errores:

Identificación de Problemas

- **Valores negativos:** Encuentras transacciones con montos negativos, lo cual es inesperado. Estos valores podrían ser errores de entrada o indicar algún problema en los datos. En este caso, puedes decidir que los montos negativos son inválidos y deben ser corregidos o eliminados.

Identificación de Problemas

- **Valores extremadamente altos o bajos:** Al revisar los montos, te das cuenta de que algunos valores son demasiado altos o demasiado bajos en comparación con el rango esperado para las transacciones financieras. Estos pueden ser valores atípicos o errores de entrada. Para abordar esto, puedes aplicar técnicas estadísticas para detectar y manejar los valores atípicos, como eliminarlos o reemplazarlos por valores más razonables.

Identificación de Problemas

- **Errores de formato:** Algunos montos pueden tener errores de formato, como la presencia de caracteres no numéricos o símbolos incorrectos. Estos errores de formato pueden dificultar los cálculos posteriores o el análisis de los datos. En este caso, puedes utilizar técnicas de limpieza de datos para corregir los errores de formato, eliminando los caracteres no deseados o reemplazándolos por valores adecuados.

Evaluación de la Calidad de los Datos

Aquí se analiza la calidad de los datos y se determina si cumplen con los estándares establecidos. Se pueden utilizar métricas como la integridad, la precisión y la coherencia para evaluar la calidad.

- Imaginemos que estás trabajando con un conjunto de datos que contiene información demográfica de una ciudad.
- Durante la fase de evaluación de la calidad de los datos, puedes realizar las siguientes verificaciones:
  - **Integridad:** Verificar que todos los campos obligatorios estén presentes y no contengan valores faltantes. Por ejemplo, puedes revisar si hay registros sin información en campos como "nombre", "edad" o "dirección". Si encuentras campos obligatorios incompletos, puedes considerar opciones como eliminar esos registros o buscar formas de completar la información faltante.

Evaluación de la Calidad de los Datos...

- **Precisión:** Evaluar la precisión de los datos verificando si los valores están dentro de rangos válidos. Por ejemplo, si la columna "edad" solo puede contener valores entre 0 y 100, puedes buscar valores que estén fuera de ese rango. Si encuentras valores incorrectos, puedes corregirlos o eliminarlos según corresponda.
- **Consistencia:** Comprobar que los datos sean consistentes en todo el conjunto. Esto implica buscar discrepancias o contradicciones entre diferentes campos. Por ejemplo, si hay una columna "género" y otra columna "título" que indica "Sr." o "Sra.", puedes verificar que no haya registros con una combinación inconsistente, como "género = Masculino" y "título = Sra.". En caso de encontrar inconsistencias, deberás tomar medidas para corregirlas o aclararlas.

Evaluación de la Calidad de los Datos...

Estas son solo algunas verificaciones que se pueden realizar durante la evaluación de la calidad de los datos. Dependiendo del contexto y los requisitos del proyecto, se pueden aplicar otras métricas y técnicas para evaluar la calidad de los datos, como verificaciones de coherencia referencial, comprobación de formatos o validación de reglas de negocio específicas.

Tratamiento de Valores Faltantes

Los valores faltantes pueden ser problemáticos, ya que pueden afectar los análisis y los modelos. En esta fase, se pueden aplicar técnicas como el relleno de valores, la eliminación de registros o la estimación basada en otros atributos para tratar los valores faltantes.

- Supongamos que estás trabajando con un conjunto de datos que contiene información sobre empleados de una empresa, y uno de los campos es "Salario".
- Durante la fase de tratamiento de valores faltantes, puedes aplicar las siguientes técnicas:

Tratamiento de Valores Faltantes...

- **Eliminación de registros:** Si el número de registros con valores faltantes es relativamente pequeño en comparación con el tamaño total del conjunto de datos y no son críticos para el análisis que realizarás, puedes optar por eliminar esos registros por completo. Por ejemplo, si solo el 5% de los registros tienen valores faltantes en la columna "Salario" y tienes suficientes datos restantes para tu análisis, puedes eliminar esos registros sin afectar significativamente la calidad de los datos.

Tratamiento de Valores Faltantes...

- **Relleno con valores estadísticos:** Si la columna "Salario" tiene valores faltantes y es importante para tu análisis, puedes utilizar valores estadísticos como la media, la mediana o la moda para rellenar los valores faltantes. Por ejemplo, puedes calcular la mediana de los salarios existentes y utilizar ese valor para completar los registros con valores faltantes. Esto puede ayudar a mantener la distribución general de los salarios y evitar sesgos en el análisis.

Tratamiento de Valores Faltantes...

- **Estimación basada en otros atributos:** Si tienes información adicional en el conjunto de datos que está correlacionada con el salario, puedes utilizarla para estimar los valores faltantes. Por ejemplo, si tienes la columna "Experiencia laboral" y encuentras una correlación positiva entre la experiencia y el salario, puedes utilizar un modelo de regresión para predecir los salarios faltantes en función de la experiencia laboral.

Tratamiento de Valores Faltantes...

- **Categorización de valores faltantes:** En algunos casos, los valores faltantes pueden tener un significado específico. Por ejemplo, si el salario falta porque un empleado es contratista y no tiene un salario fijo, puedes asignar una categoría especial como "Contratista" en lugar de un valor numérico. Esto ayuda a mantener la integridad y la coherencia de los datos.

Eliminación de Valores Atípicos

Los valores atípicos son valores que difieren significativamente del resto de los datos. Pueden ser errores o indicar situaciones excepcionales. En esta etapa, se pueden utilizar técnicas estadísticas para detectar y manejar los valores atípicos, como la eliminación o el reemplazo por valores más apropiados.

Eliminación de Valores Atípicos...

- Identificación de valores atípicos: Utiliza métodos estadísticos, como el cálculo de los límites de valores atípicos utilizando la desviación estándar o los percentiles, para identificar los valores que están significativamente fuera del rango esperado. Por ejemplo, si la mayoría de las calificaciones se encuentran entre 0 y 100, pero encuentras algunas calificaciones extremadamente altas, como 1000, es probable que sean valores atípicos.

Eliminación de Valores Atípicos...

- Visualización de datos: Utiliza gráficos, como histogramas o diagramas de caja, para visualizar la distribución de las calificaciones y detectar visualmente los valores atípicos. Los valores que estén muy por encima o por debajo de los valores esperados pueden considerarse atípicos.

Eliminación de Valores Atípicos...

- Eliminación de valores atípicos: Una vez identificados los valores atípicos, puedes decidir eliminarlos del conjunto de datos. Esto implica eliminar los registros completos que contienen esos valores o reemplazar los valores atípicos por valores faltantes, para tratarlos posteriormente en el proceso de imputación de datos faltantes. La eliminación de valores atípicos debe realizarse con precaución, ya que puede afectar la representatividad de los datos y el análisis posterior.

Eliminación de Valores Atípicos...

Es importante tener en cuenta que la eliminación de valores atípicos debe realizarse de manera cuidadosa y considerando el contexto del problema y los requisitos del análisis. Los valores atípicos pueden surgir debido a errores de entrada, mediciones incorrectas o casos genuinos de datos inusuales, y su eliminación puede tener un impacto en la interpretación y validez de los resultados. Por lo tanto, es esencial evaluar cuidadosamente la relevancia de los valores atípicos y considerar si deben ser eliminados o si requieren un tratamiento especial en función del conocimiento del dominio y los objetivos del análisis.

Corrección de errores de formato y estandarización

Aquí se abordan problemas relacionados con el formato de los datos, como errores de entrada, inconsistencias en las unidades de medida o problemas de codificación. Se realizan transformaciones para asegurar que los datos sigan un formato coherente y estandarizado.

Supongamos que estás trabajando con un conjunto de datos que contiene información de direcciones de clientes. Durante la fase de corrección de errores de formato y estandarización, puedes aplicar las siguientes técnicas:

Corrección de errores de formato y estandarización...

Corrección de formato: Verifica si las direcciones siguen un formato consistente y correcto. Por ejemplo, puede haber errores como letras mayúsculas o minúsculas incorrectas, espacios adicionales, caracteres especiales no válidos o inconsistencias en la estructura de la dirección. Puedes utilizar técnicas de limpieza de datos para corregir estos errores, como convertir todo el texto a mayúsculas o minúsculas, eliminar espacios adicionales o reemplazar caracteres especiales incorrectos.

Corrección de errores de formato y estandarización...

Normalización de direcciones: Si las direcciones están en diferentes formatos o siguen diferentes convenciones, puedes estandarizarlas a un formato común. Por ejemplo, puedes utilizar librerías o servicios de geocodificación para estandarizar las direcciones a un formato aceptado por los servicios de geolocalización. Esto puede implicar la normalización de elementos como la abreviatura de la calle, el uso de números cardinales en lugar de ordinales, la estandarización de nombres de calles o la inclusión de códigos postales en un formato específico.

Corrección de errores de formato y estandarización...

Validación de direcciones: Además de corregir el formato, puedes realizar validaciones de direcciones para asegurarte de que sean direcciones reales y válidas. Puedes utilizar servicios de validación de direcciones que verifiquen si la dirección existe en una base de datos oficial o aplicar reglas de validación específicas según las convenciones de tu país o región. Esto ayuda a garantizar la calidad y precisión de los datos de dirección.

Corrección de errores de formato y estandarización...

Enriquecimiento de datos: Si los datos de dirección son incompletos o contienen errores significativos, puedes utilizar fuentes externas o servicios para enriquecer los datos. Por ejemplo, puedes utilizar servicios de geocodificación para agregar información adicional a partir de la dirección, como coordenadas geográficas o información de ubicación precisa.

Corrección de errores de formato y estandarización...

Es importante tener en cuenta que la corrección de errores de formato y estandarización puede ser un proceso complejo y depende del contexto y los requisitos del proyecto. Además, es fundamental tener un conocimiento sólido de las convenciones de formato y validación aplicables a las direcciones en tu área geográfica específica. La corrección y estandarización adecuadas de los datos de dirección mejoran la coherencia, la calidad y la utilidad de los datos, facilitando su posterior análisis y aplicación.

Verificación de la coherencia y la integridad

En esta fase, se realizan comprobaciones para asegurarse de que los datos sean coherentes y estén libres de duplicados. Se pueden utilizar técnicas como la verificación de integridad referencial, la reconciliación de datos o la detección de duplicados.

El éxito de la limpieza de datos se puede evaluar en función de diferentes criterios:

- **Calidad mejorada:** Los datos limpios deben tener una calidad superior en comparación con los datos originales, lo que significa que deben ser más precisos, completos y coherentes.
- **Mayor confiabilidad:** Los datos limpios son más confiables para su uso en análisis y toma de decisiones, ya que se han abordado los errores y se ha mejorado su calidad.

- **Resultados más precisos:** Al limpiar los datos, se reducen los sesgos y las distorsiones que podrían afectar los resultados de los análisis o los modelos, lo que lleva a resultados más precisos y confiables.
- **Eficiencia mejorada:** Con datos limpios, los científicos de datos pueden ahorrar tiempo y esfuerzo al evitar errores costosos y retrabajos.