Profundiza más

Recurso de Profundización

Ideas necesarias para la limpieza de datos

Introducción

La limpieza de datos es un paso necesario en cualquier procesamiento de datos, de lo que se trata es de eliminar errores, valores atípicos e incoherencias, de modo que garantiza que los datos estén en un formato adecuado para el análisis y que no se produzcan alteraciones en el proceso. En ese sentido, la limpieza de datos permite obtener resultados precisos y fiables, que puedan llevar a conclusiones reales.

Los datos sin limpiar, se llaman datos sucios, estos pueden hacer referencia a

- ⇒ **Valores perdidos**: Datos incompletos que pueden sesgar los resultados.
- ⇒ **Valores atípicos**: Valores fuera de la norma que pueden sesgar los resultados.
- ⇒ **Duplicados**: Entradas repetidas que pueden sobre representar una entrada en el análisis.
- ⇒ **Datos erróneos**: Valores incorrectos que pueden ser difíciles de identificar.
- ⇒ **Incoherencias**: Datos inconsistentes en formato o unidades.

Técnicas de limpieza de datos

- ⇒ **Tratamiento de valores perdidos**:
  - Eliminar entradas con valores perdidos, usualmente se pierde el dato de ese participante.
  - Imputar valores razonables, a través del cálculo de la media y la mediana se agrega un valor medio en donde falta el dato en la base, de modo que no altere los resultados pero que no se pierdan los otros datos recolectados de ese participante.
- ⇒ **Detección y tratamiento de valores atípicos**:
  - Identificar valores atípicos mediante los diferentes valores obtenidos por la estadística descriptiva, a veces los métodos gráficos facilitan el localizarlos.
  - Tratar valores atípicos corrigiendo errores, eliminándolos o sustituyéndolos.
- ⇒ **Tratamiento de duplicados**:
  - Identificar filas duplicadas y eliminarlas.
  - Fusionar registros duplicados agregando información.
- ⇒ **Abordar las incoherencias**:

Profundiza más

- Corregir errores y erratas, a veces, es necesario revisar los registros originales de la recolección de datos para determinar de donde parte la incoherencia,
- Convertir las variables cualitativas a cuantitativas, de modo que a través de una categorización numérica puedan procesarse. Esto quiere decir a que a los datos cualitativos se les asigna un número, por ejemplo, en un base de datos donde las personas marcaron su género como masculino o femenino, se puede asignar el valor 1 a masculino y el valor 2 a femenino, así podremos relacionar variables cuantitativas y cualitativas.

Exploración y preprocesamiento de datos

La exploración de datos hace referencia al proceso de comprender el conjunto de datos antes del análisis, el investigador debe conocer su base datos de modo que, si encuentra alguna inconsistencia, pueda reconocer con facilidad de qué se trata para resolverla.

Después de ello, se realiza el preprocesamiento de datos, en donde se examina la fuente de los datos (actualmente, gran parte de los investigadores, los recolectan a través de formularios en línea, lo que facilita su consolidación en bases de datos) y su contexto. De modo que se sabe cual es el origen el método de recolección de datos. Esto debe detallarse en la metodología, para indicar la confibialidad del proceso de recolección.

Mas adelante, se determina el número de variables a estudiarse y sus categorías, tal como se vio en el desarrollo de la clase. Determinar el número y la categoría de las variables nos permitirá comprender qué prodedimientos estadísticos deben realizarse y entre qué variables.

Luego de ello, se cargan los datos en el programa elegido para su procesamiento y se realizan los procedimientos estadísticos correspondientes, suele empezarse con la estadística descriptiva y se generan los gráficos que indiquen los datos de una manera sintética. Además, la misma estadística descriptiva a través de la desviación estándar y la varianza, dan indicios de si hay diferencias entre los datos que puedan ser indagadas a través de las pruebas de hipótesis.

Recomendaciones

- ⇒ **Una vez realizada la limpieza de datos, es importante que se mantengan los registros originales por si ocurre algún imprevisto, y además, se constituyen como evidencia de la recolección de datos.**
- ⇒ **Se debe guardar la base con los datos sucios, pues si aparecen inconsistencias durante el procesamiento, se puede volver a ella y revisar.**

Profundiza más

- ⇒ **Almacer un documento de respaldo con la base de datos limpia, así como, documentar qué datos fueron parte del proceso de limpieza de datos.**
- ⇒ **Realiza los procedimientos de manera adecuada, evita introducir datos aleatorios para no crear sesgo en el estudio.**

Importancia:

La importancia de la limpieza de datos radica en que nos permite comprender la composición de nuestros datos, y por lo tanto, los límites que tiene. Así podemos agilizar los procedimientos estadísticos, pues tendremos claridad sobre aquellos que podemos realizar y aquellos que no son oportunos.

Además, garantiza que los datos sean precisos, coherentes y confiables. Por ello, nuestros resultados podrán exponerse sin complicaciones.