Análisis exploratorio y preparación de los datos

Año 2014

La empresa CarMD, dedicada a la venta de autos usados, ha recolectado información sobre el consumo [l/100km], aceleración [seg. de 0 a 100 km/h] y velocidad final [km/h] de cada uno de los vehículos que tiene en venta.

El señor Gunderson, gerente de la empresa, nos ha solicitado efectuar un análisis descriptivo de sus datos, incluyendo la limpieza de los mismos si fuera necesario (ver anexo).

1. Efectuar un análisis univariante considerando:
    a. Medidas de posición y dispersión
    b. Histograma de frecuencias
    c. Gráfico de cajas y bigotes (box plot)
    d. Análisis de normalidad con QQ-plot y test de Shapiro Wilk
    e. Análisis de curtosis

2. Efectuar un análisis multivariante evaluando:
    a. Vector de medias, matriz S y matriz R.
    b. Gráficos de dispersión

Gunderson nos solicitó que tengamos en cuenta el hecho de que cada vehículo pertenece a una clase diferente (estándar o premium), ya que sospecha que pueden tener características diferentes.

3. Efectuar un nuevo análisis teniendo en cuenta el tipo de vehículo.
    a. Diagrama de cajas y bigotes estratificado
    b. Diagramas de dispersión estratificados
    c. Gráficos en 3D

    ¿Tiene relación la clase de vehículo con sus características?

Analizando los datos, usted recordó que según las nuevas normas en consumo de combustible (ArgenOIL), un auto se considera ecológico cuando consume menos de 7 lts. de combustible cada 100 km.

4. Construya una nueva variable que diferencie los autos ecológicos de los que no lo son.

5. Analice las relaciones de esta nueva variable con las restantes.

A lo largo del proceso usted ha detectado y resuelto algunos problemas en la calidad de los datos, por la presencia de valores atípicos (outliers) y datos faltantes (NA).

6. Realice un informe de calidad de los datos, donde se describan las actividades de limpieza que se hayan efectuado.

Métodos para la limpieza de datos

Detección de outliers

Detección univariante de casos atípicos
- Gráfico de caja y bigotes
- Estadísticos robustos de la variable

Detección bivariante de casos atípicos
- Gráfico de caja y bigotes
- Gráficos de dispersión

Detección multivariante de casos atípicos
- Estadísticos basados en distancia
- Análisis de componentes principales

Métodos para la solución de datos faltantes

Bajo el supuesto de que los datos faltantes (NA, Not Available) se distribuyen aleatoriamente en la muestra, es posible utilizar métodos de supresión o imputación:

- Supresión de datos
  - Aproximación de casos completos o supresión de casos según lista
  - Suprimir casos (filas) o variables (columnas)

- Imputación de la información faltante
  - Imputación por sustitución del caso
  - Imputación de sustitución por la media
  - Imputación de sustitución por la mediana
  - Imputación por regresión