Pecotche Andres
Lucas Carballo

Minería de Datos usando Sistemas Inteligentes

PRACTICA 1 – PREPROCESAMIENTO DE LOS DATOS

a) *Discretizar por frecuencia el atributo Edad.*

Primeramente ordenamos los valores del atributo edad, quedando el orden de los valores de la siguiente manera:

-300, -155, -60, -32, -25, -10, -7,-1, 2, 10, 14, 16, 120

Cantidad de elementos: **13**. → Impar (6,7). (6 elementos “bajos” y 7 “altos)
N=2 → (Dos intervalos: **Bajo** y **alto**).

- BAJO → **-300, -155, -60, -32, -25, -10**
- ALTO → **-7,-1, 2, 10, 14, 16,120**

Separador entre intervalos: (-10-7) / 2 = **-8.5**

	BAJA	ALTA
INTERVALO	[-∞,-8.5)	[-8.5, ∞)
VALORES	-300, -155, -60, -32, -25, -10	-7,-1, 2, 10, 14, 16,120

b) *Discretizar por Rango el Atributo Edad.*

Longitud del rango = 120-(-300)= **420**.
Largo de valores = 420/n → 420/2 = **210**.
Valores intermedios= -300+210= **-90**.

	BAJA	ALTA
INTERVALO	(-∞,-90)	[-90, ∞)
VALORES	-300, -155	-60, -32, -25, -10, -7,-1, 2, 10, 14, 16,120

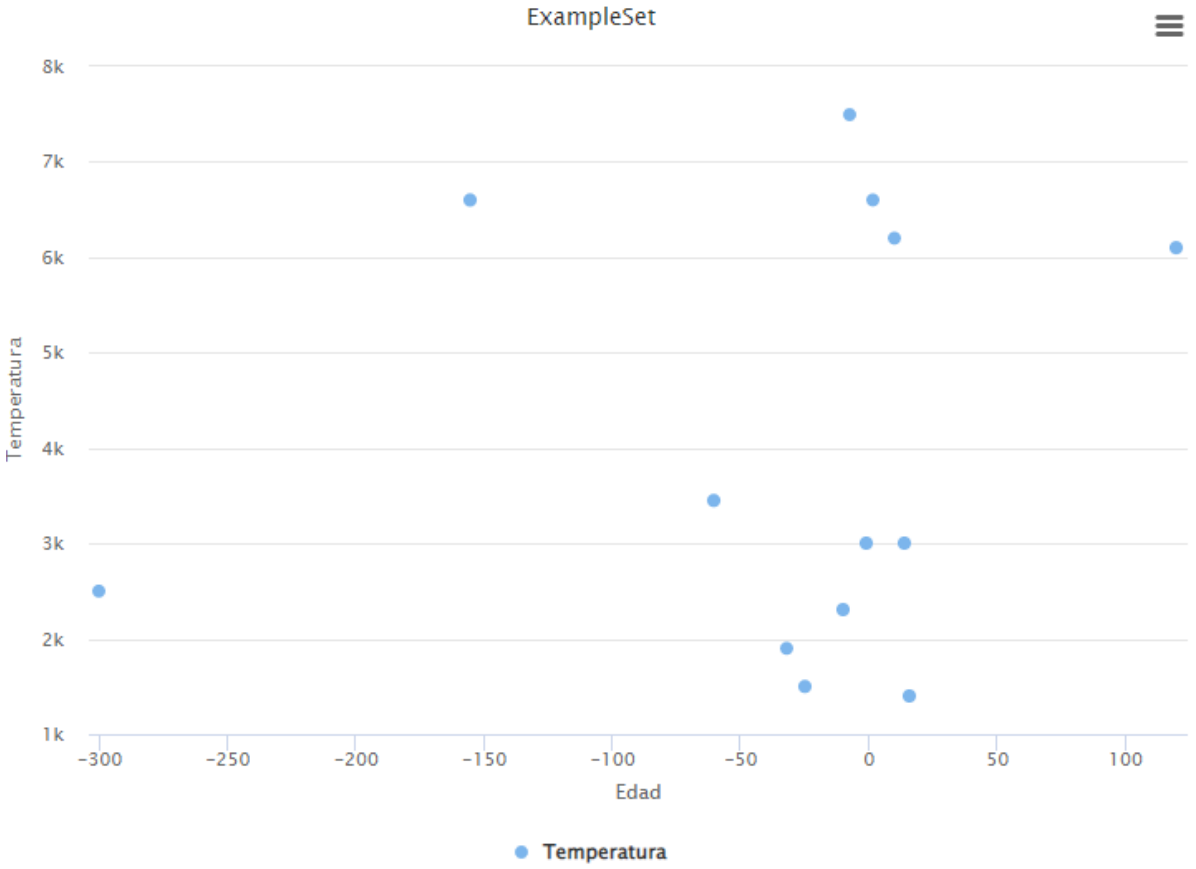
c) *Correlación lineal entre Edad y Temperatura.*

X = edad, Y = temperatura

- Índice de correlación (x,y) = (Cov (x,y)) / (desv (x) * desv (y))
- Cov(x,y) = $\frac{[\sum_{i=1}^{N}(x_i-\mu_x)*(y_i-\mu_y)]}{N-1}$ = **35703.85**
- desv (x) = $\sigma_x = \frac{\sqrt{\sum_{i=1}^{N}(x_i-\mu_x)^2}}{N}$ = **96.19**

- desv (y) = $\sigma_y = \frac{\sqrt{\sum_{i=1}^{N}(y_i - \mu_y)^2}}{N}$ = **2146.23**

Valor Índice de correlación = 35703,85 / (96,19 * 2146,23) = **0.175**
Intensidad: Débil (<0,5) (no hay correlación lineal)
Tipo: Covarianza cercana a cero. En el siguiente gráfico se puede apreciar esto.



Valor	0.175
Intensidad	Débil
Tipo	Covarianza cercana a cero.

d) Diagrama de caja de tukey de la variable edad.

- Mediana:
  N=13 →Impar.
  x̄= -7
- Q1:
  Ubicación = (N+1)/4 → (13+1)/4= **3.5**
  Q1= X_3+0.5(X_4-X_3) → -60+0.5(-32+60)= **-46**

- Q3:
  Ubicación =3(N+1)/4 → 3(13+1)/4= **10.5**
  $Q3=X_{10}+0.5(X_{11}-X_{10})$ → 10+0.5(14-10)= **12**

- RIC:
  RIC= Q3 - Q1 = 12 - (-46) = 12 + 46 = **58**

- Bigote Superior:
  Q3+1.5*RIC → 12+1.5*58=**99** → El dato anterior a este valor nos va a dar el bigote superior, por lo tanto, el bigote superior es **16**.

- Bigote Inferior:
  Q1-1.5*RIC → -46-1.5*58= **-133** → El dato anterior a este valor nos da el bigote inferior, por lo tanto, el bigote inferior es **-60**.

- Intervalos de valores atípicos leves

  (Q1– 3*RIC, Q1 – 1.5*RIC) y (Q3 + 1.5*RIC, Q3 + 3*RIC) =

  Intervalo inferior: (-220, -133)
  Intervalo superior: (99, 186)

- Valores atípicos leves

  Intervalo inferior: [-155]
  Intervalo superior: [120]

- Intervalos de valores atípicos extremos

  (-inf, Q1– 3*RIC) y (Q3 + 3*RIC, +inf) =
  Intervalo inferior: (-inf, -220)
  Intervalo superior: (186, +inf)

- Valores atípicos extremos

  Intervalo inferior: [-300]
  Intervalo superior: No se encuentran valores atípicos extremos superiores en la muestra.

Mediana	-7

Q1	-46
Q3	12
RI	58
Bigote superior	16
Bigote inferior	-60
Intervalos de valores atípicos leves	(-220, -133) y (99, 186)
Valores atípicos leves	-155, 120
Intervalos de valores atípicos extremos	(-inf, -220) y (186, +inf)
Valores atípicos extremos	-300