Electiva Aprendizaje automático
Análisis exploratorio de datos

Maestría en Gestión de Tecnologías de Información y Conocimiento
Facultad de Ingeniería
Universidad de Nariño

MSc Jimmy Mateo Guerrero Restrepo[1]
jimaguere@udenar.edu.co
3168211698

[1]Departamento de Sistemas
Facultad de Ingeniería

Aprendizaje automático, 2021

Contenido

Análisis Univariado

- El análisis descriptivo es el paso inicial generalmente recomendado para comprender la estructura de los datos disponibles y la extracción de la información relevante para el análisis.
- Las variables pueden clasificarse según su nivel de medición:
  - Categóricas o cualitativas: No se puede establecer un ordenamiento entre ellos, ejemplo color de cabello, sexo, profesión del padre, marca de desodorante, etc.
  - Cuasicuantitativas u Ordinales: Se puede establecer un ordenamiento entre las categorías, no se puede establecer una distancia entre los distintos niveles de las mismas. Calificación de examen (A, B, C, D y E), orden de mérito en un concurso, etc.
  - Cuantitativas Discretas: Se vinculan generalmente con el proceso de contar. cantidad de hijos, cantidad de materias aprobadas, dinero disponible en una billetera, etc.
  - Cuantitativas Continuas: Se asocian generalmente al proceso de medir. Ejemplos el peso, edad, duración de un llamado, etc.

Para datos cualitativos

El paso inicial más sencillo es confeccionar una tabla denominada distribución de frecuencias; que tiene un aspecto particular para cada tipo de variable de las consideradas.

- Las clases se definen según el interés de la investigación.
- Se cuenta la cantidad de observaciones de cada clase.

Para datos cualitativos

Ejemplo: estudiamos los modelos de autos vendidos en una sucursal de una concesionaria de autos Chevrolet.

modelo	Nro. de observaciones
Tracker	6
Spin	10
Cruze	7
Agile	12
Onix	17
Clasic	32
Sonic	18
Celta	25

Para datos cuantitativos

- Discretas las clases quedan definidas por los valores de la variable.
- CONTINUAS es necesario crear intervalos de clase.
- En ambos casos, se registra la cantidad de observaciones de cada categoría y a dicha cantidad se la denomina frecuencia absoluta.

Para datos cuantitativos discretos

Ejemplo: Estudiamos ahora la evolución de las ventas de vehículos de alta gama, en la misma sucursal durante los últimos 24 meses.

Cantidad de vehículos de alta gama vendidos	Cantidad de meses
1	2
2	3
3	7
4	4
5	8
7	1

Para datos cuantitativos continuos

Ejemplo: estamos interesados en investigar la cantidad de proteínas consumidas por día para una muestra de habitantes de distintas zonas de Colombia.

Intervalos de clase	f_i (frec. absoluta)
[7; 9)	6
[9; 11)	10
[11; 13)	4
[13; 15)	7
[15; 17)	5

Esta tabla informa, por ejemplo, que 4 individuos consumieron entre 11 y 13 g de proteínas por día.

Outline

Medidas resumen univariadas

Entre las más populares, podemos mencionar la media, la moda, mediana y los cuartiles.

- Media aritmética: frecuentemente denominada promedio de los datos obtenidos de la muestra. En su cálculo intervienen todos los valores observados.

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

es muy influenciada por la presencia de valores extremos.

Medidas resumen univariadas

Mediana:

- Es el valor central de los datos ordenados de menor a mayor. Deja al 50 % de las observaciones por debajo de sí y al 50 % de las observaciones por encima.
- La Mediana depende de la posición que ocupan los datos en la ordenación y no del valor numérico de los mismos.

Medidas resumen univariadas

- Si la distribución de la variable en estudio no es simétrica la media y la mediana no coinciden:

$\bar{X}$ > Me ⟶ asimetría por la derecha o positiva

$\bar{X}$ < Me ⟶ asimetría por la izquierda o negativa.

a- 12, 13, **15**, 18, 23 Md = 15

b- 12, 13, **15**, **18**, 23, 35 Md= (15+18)/2

Si el tamaño de la muestra o población de estudio es mayor:

Para variables cuantitativas discretas: si recorremos la tabla de frecuencias por la columna de las acumuladas en orden ascendente, la mediana es el primer valor de variable que registra una frecuencia acumulada superior o igual a la mitad de las observaciones.

x_i	f_i	F_i
1	10	10
2	15	25
3	20	45
4	5	50

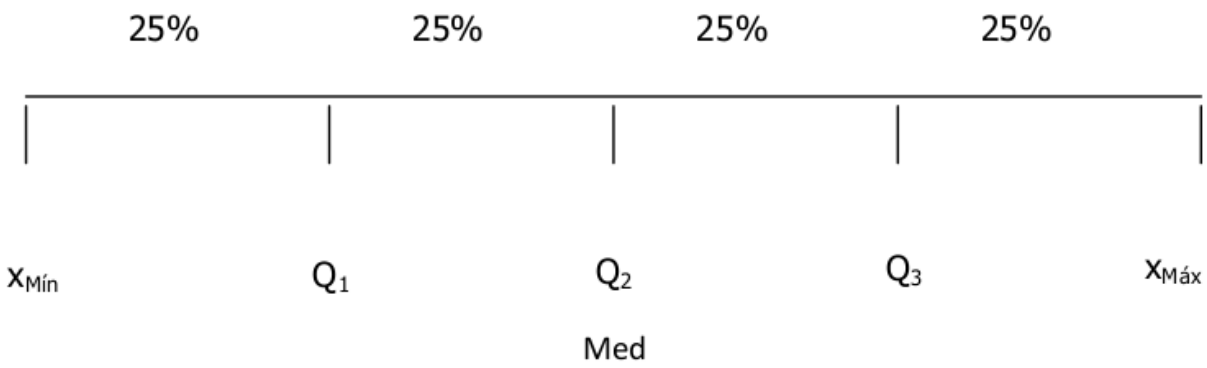
Md = (2+3)/2= 2,5

x_i	f_i	F_i
1	10	10
2	14	24
3	21	45
4	5	50

Md= 3

Medidas resumen univariadas

- Modo o Moda: es el valor observado de mayor frecuencia. Un mismo grupo de datos puede tener más de un Modo. Se representa con Mo y tiene las mismas unidades de medida que la variable original.
- Cuartiles: son los valores de la variable, observados o no, que dividen a la muestra en cuatro partes iguales, o sea, que dividen a la muestra en cuartos. Entonces tendremos en la muestra tres cuartiles:



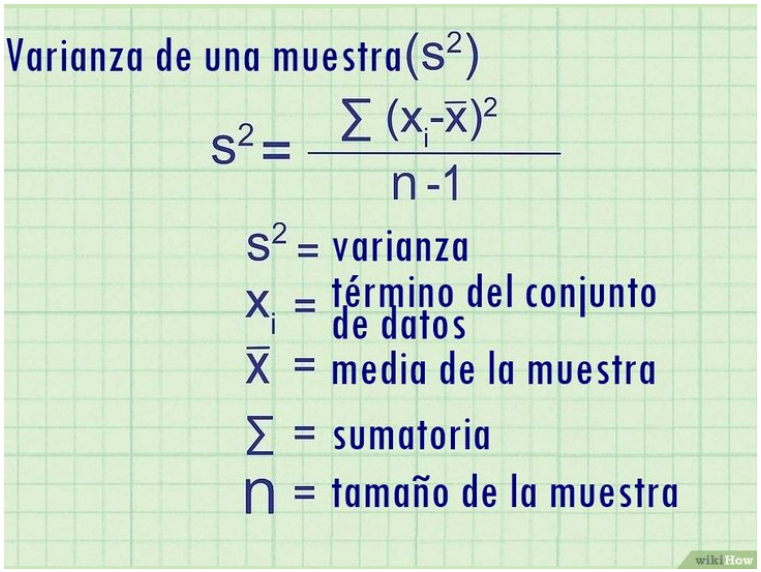
Outline

Medidas de dispersión

Miden el grado de concentración de los datos respecto de alguna medida resumen de centralidad.

- Rango o Amplitud: Es la diferencia entre el máximo valor observado y el mínimo observado.
- Varianza muestral:Se expresa en unidades de la variable al cuadrado. Dado un conjunto de n valores, x1 , x2 , ..., xn , la varianza muestral es la suma de los cuadrados de los n desvíos entre cada valor Xi y la media aritmética de los mismos, dividida por (n –1).

Medidas de dispersión



Mide el Grado de concentración de las observaciones con respecto a su media aritmética.

Medidas de dispersión

- Desvío estándar muestral: es la raíz cuadrada positiva de la varianza muestral. Se representa con S o Sx. Se expresa en la misma unidad que la variable.
- Coeficiente de Variación (CV): es el cociente entre el desvío estándar muestral y la media aritmética. Se suele expresar en porcentaje.

$$CV = \frac{S_x}{|\bar{x}|}$$

S_x = Desviación típica del conjunto de datos.

$|\bar{x}|$ = Valor absoluto de la media del conjunto de datos $(X_1, X_2, \ldots, X_n)$ y $\bar{x} \neq 0$.

Indica el porcentaje que la desviación estándar representa de la media aritmética.

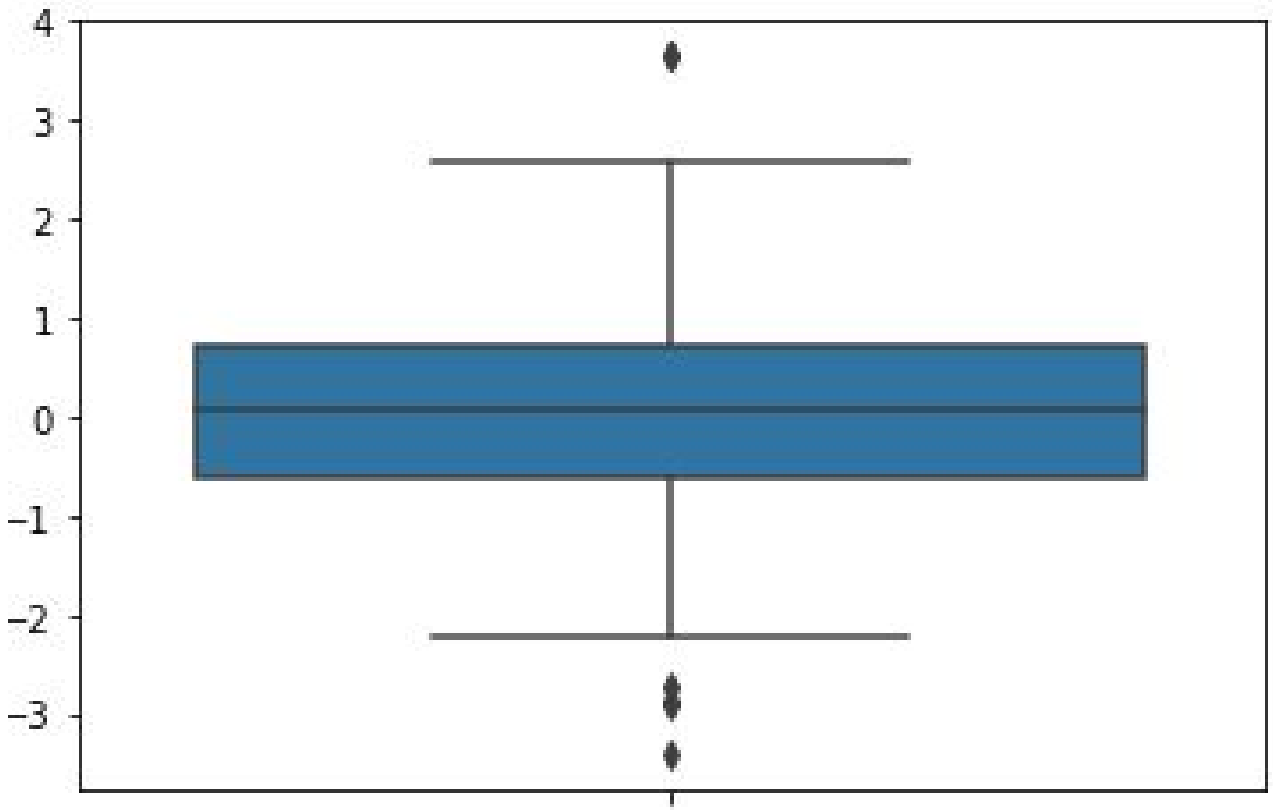
Medidas de dispersión

Distancia intercuartílica (DI): mide la dispersión de los valores centrales. La forma de calcularla es:

- DI = Q3 - Q1 , e indica el rango donde se encuentra el 50 % central de las observaciones.
- La distancia intercuartílica indica el rango que abarca el 50 % central de las observaciones.

Boxplot o Gráfico de Caja y Bigote (Box and Whisker Plot)

Tukey (1977) propuso este gráfico para presentar datos numéricos, apreciar características importantes de la distribución y comparar distintas distribuciones. Está basado en las medidas de posición. Es un gráfico de fácil lectura.



Boxplot o Gráfico de Caja y Bigote (Box and Whisker Plot)

a. Se dibuja una caja (box) cuyos extremos son los cuartiles primero y tercero y, dentro de ella, un segmento que corresponde a la mediana o segundo cuartil.

b. A partir de cada extremo dibujamos un segmento (bigote), hasta el dato más alejado que está, a lo sumo, a 1,5 veces DI del extremo de la caja.

c. Se marca con ▪ a aquellos datos que están entre 1,5 y 3 veces DI de cada extremo de la caja (outlier moderado) y con * a aquellos que están a más de 3 veces DI de cada extremo de la caja (outlier severo).

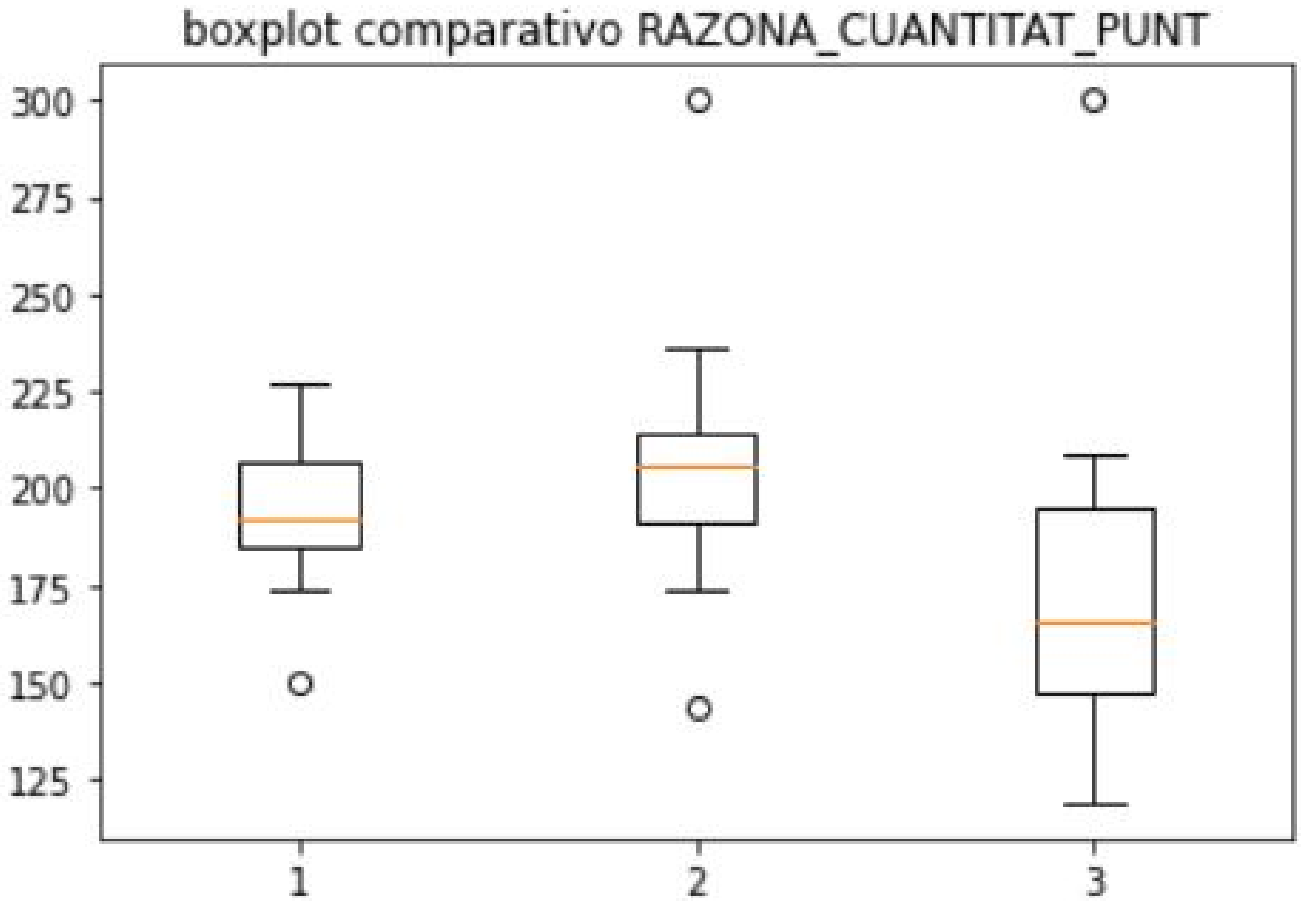
Boxplot o Gráfico de Caja y Bigote (Box and Whisker Plot)

A partir de un boxplot podemos apreciar los siguientes aspectos de la distribución de un conjunto de datos:

- Posición
- Dispersión
- Asimetría
- Puntos anómalos o outliers.

Boxplot o Gráfico de Caja y Bigote (Box and Whisker Plot)

Los box-plots son especialmente útiles para comparar varios conjuntos de datos, pues nos dan una rápida impresión visual de sus características.




MEAN
MEDIAN
OUTLIER
imgflip.com

Datos atípicos, salvajes (outliers)

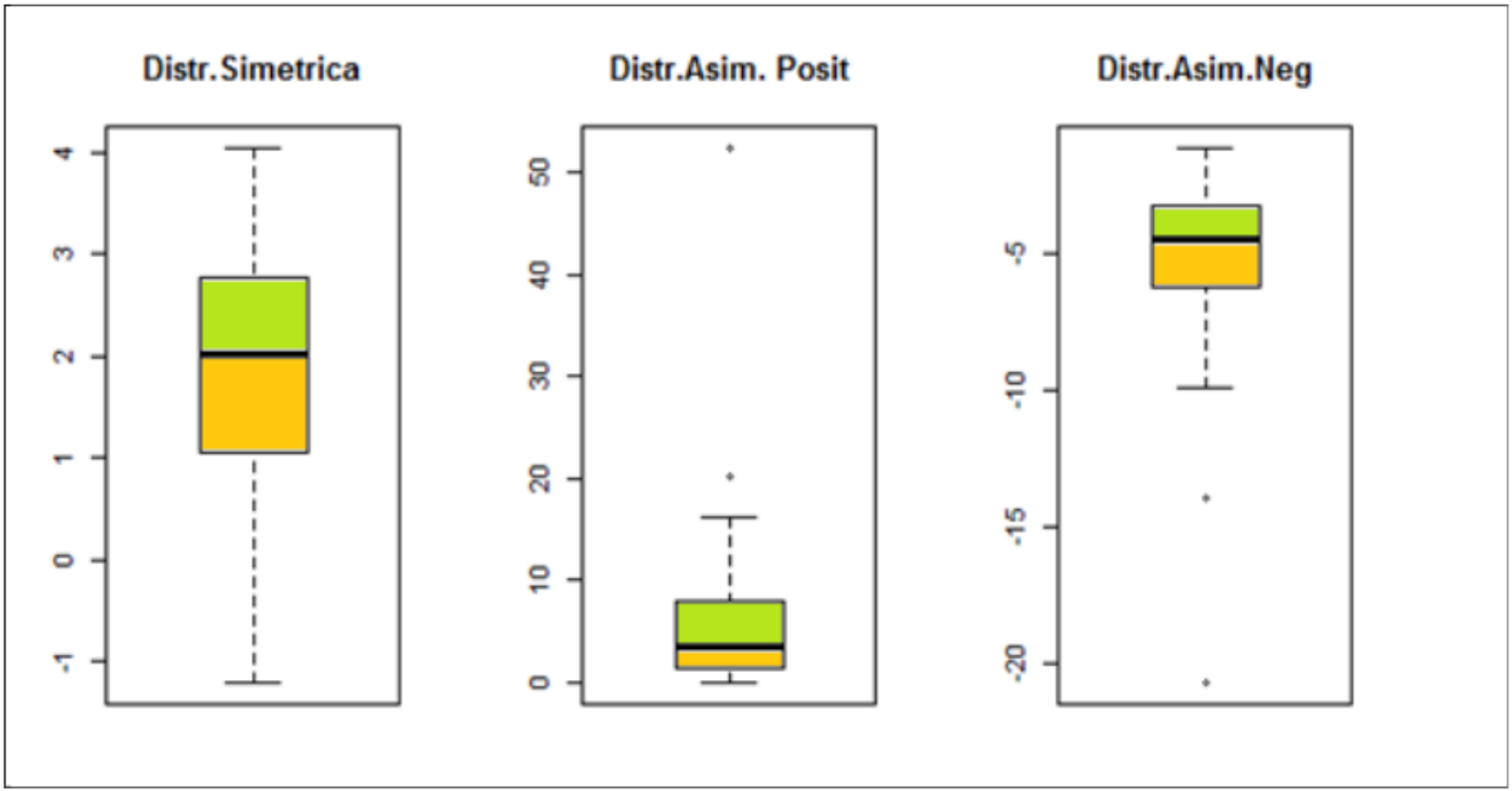
- La detección de observaciones atípicas es importante pues pueden determinar o influenciar fuertemente los resultados de un análisis estadístico clásico. Esto ocurre porque muchas de las técnicas habitualmente usadas son muy sensibles a la presencia de estos datos.
- Los outliers deben ser cuidadosamente inspeccionados, si no hay evidencia de error y su valor es posible no deben ser eliminados. Pueden estar alertando de anormalidades de un tratamiento o patología.
- La presencia de outliers puede indicar que la escala elegida no es la más adecuada, podemos tener una idea de cuán influyentes son los datos , en función de su alejamiento del conjunto general de datos.

Datos atípicos, salvajes (outliers)

Ejemplo : si en una muestra con n $= 13$, tenemos los siguientes datos:

- 14 18 24 26 35 39 43 45 56 62 68 92 198
- Se observa claramente que el valor 198 está alejado del grupo de valores restantes, por lo que 198 es un valor atípico (outlier)

Boxplot o Gráfico de Caja y Bigote (Box and Whisker Plot)



$\bar{X}$> Me ⟶ asimetría por la derecha o positiva

$\bar{X}$< Me ⟶ asimetría por la izquierda o negativa.

¿Cómo se presenta la información multivariada?

La forma más usual en la que se presenta un conjunto de datos multivariados es una tabla donde aparecen los valores de p variables observadas sobre n elementos.

marca	valor energético (calorías/100g)	Carbohidratos (g/100g)	Proteínas (g/100g)	grasas totales (g/100g)	sodio(mg/100g)
cerealitas	439	65	11	15	574
fajitas	466	57	10	22	828
express s/sal	445	69	11	14	12
oreo	478	67	5,6	21	363
melba	464	70	6,3	18	263
pepitos	463	66	7,1	19	136
criollitas	438	69	11	13	431
merengadas	418	69	6,3	13	201
sonrisas	423	70	6,8	13	241
maná	444	73	9	13	375
guinditas	407	70	6	12	106,7
pepas	437	60	6,7	18	76,67
polvorón	410	56,7	6,3	18	66,7
bizcoch.grasa.azuc	493	60	7,6	24	1066
hogareñas.salvado	424	65	11	13	892
granix.con.lino	462	55	11	22	931
bagley salvado	421	63	11	14	624

¿Cuáles son los objetivos del Análisis Exploratorio de Datos?

La descripción de los datos multivariantes comprende el estudio de cada variable aisladamente y también de las relaciones que quedan definidas entre ellas.

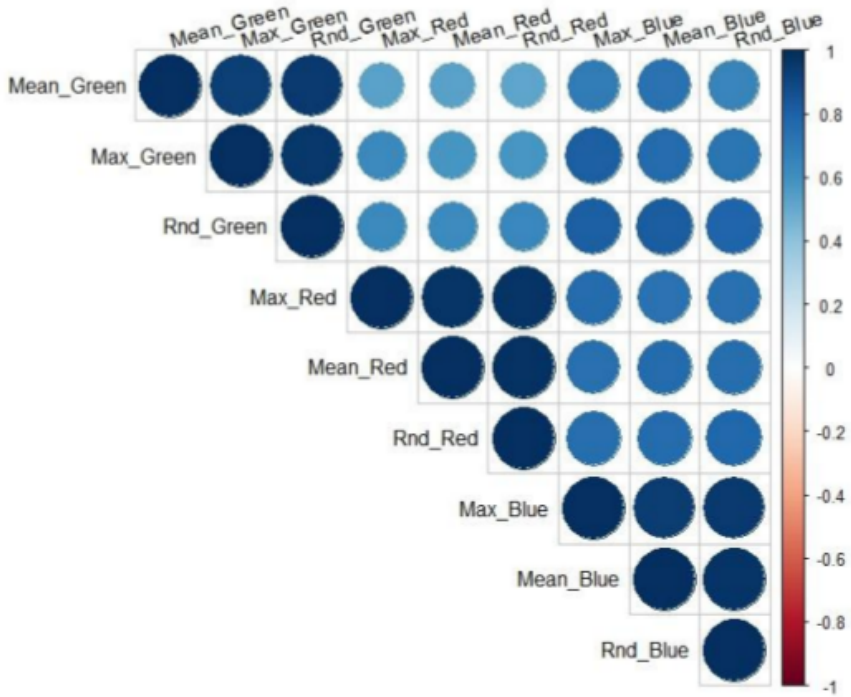
- Conocer los datos
- Descubrir patrones
- Verificar la existencia de patrones
- Resumir información
- Hallar asociaciones de variables
- Detectar anomalías

Análisis Exploratorio de Datos

Correlograma

En un correlograma (las correlaciones se miden de acuerdo al tamaño del círculo). El color rojo indica correlación negativa y el color azul correlación positiva.

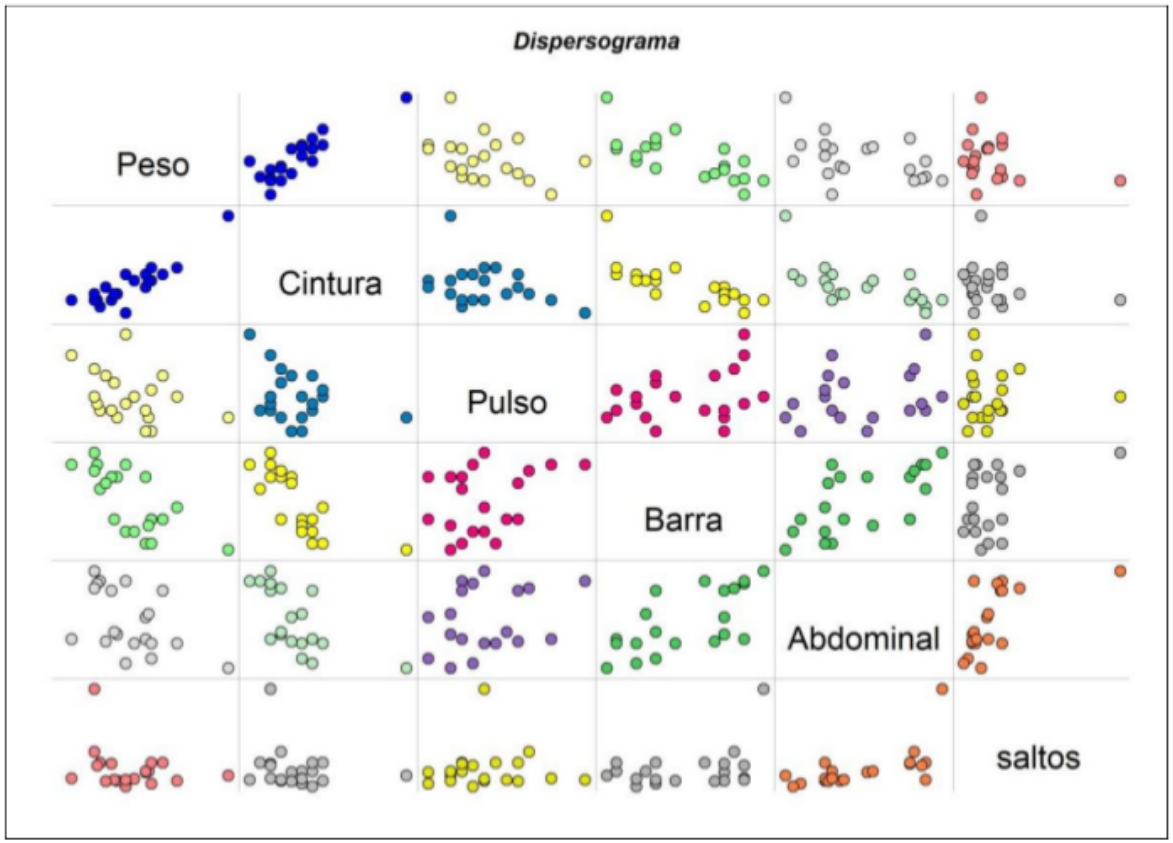
En este conjunto de variables no se aprecian correlaciones negativas.



Análisis Exploratorio de Datos

Dispersograma

En ocasiones nos interesa detectar cuales de las variables de análisis se relacionan con cuales otras y de qué forma. Para ello es útil representar todos los diagramas de dispersión de a dos.



Transformaciones del conjunto de datos

En algunas ocasiones, para optimizar el análisis de la información disponible, es conveniente realizar transformaciones a los datos. Estas transformaciones pueden ser por variable (columna) o por individuos (filas); dependiendo de los objetivos de la misma.

- Objetivos
  - Hacer comparables las magnitudes
  - Modificar la escala de medición
  - Satisfacer alguna propiedad estadística
- Dos tipos:
  - Por variable
  - Por individuo

Transformaciones por variable

Se usan para estandarizar (hacer comparables) distintas variables que serán usadas en análisis multivariados posteriores.

- Suele denominarse a esta transformación z-scores o puntuaciones Z, que tienen la característica de tener media 0 y varianza 1.

$$z_i = \frac{x_i - \mu}{\sigma} \qquad \text{para} \qquad i = 1,\ldots,n$$

Transformaciones por individuo

- Se usan para estandarizar (hacer comparables) los valores de los distintos individuos.
- Ejemplo: varios jueces evalúan un conjunto de individuos, cada juez tiene una tendencia, puntuaciones muy altas o muy bajas. . . se pueden corregir por juez los puntajes

$$\text{Si } x \geq \bar{x} \rightarrow T(x) = \frac{x - \bar{x}}{x_{max} - \bar{x}}$$

$$\text{Si } x < \bar{x} \rightarrow T(x) = \frac{x - \bar{x}}{\bar{x} - x_{min}}$$

- Con esta transformación se neutraliza la tendencia del juez.

¿Cuáles son los parámetros que debemos estimar?

- Para cada una de las variables por separado, estimamos su media y su varianza o desviación estándar.
- Para el conjunto de todas las variables, se estima el vector de medias y la matriz de varianzas y covarianzas.

El vector de medias es: $\bar{x} = \begin{pmatrix} \overline{x_1} \\ \overline{x_2} \\ \cdots \\ \overline{x_p} \end{pmatrix}$ es un vector que tienen tantas componentes como variables tenga nuestra matriz de datos. Las componentes del vector son las medias de cada una de las variables consideradas.

En nuestro ejemplo de las galletitas, el vector de medias es:

$$\bar{x} = \begin{pmatrix} \overline{valor\ energ} \\ \overline{carbohidratos} \\ \overline{proteinas} \\ \overline{grasas} \\ \overline{sodio} \end{pmatrix} = \begin{pmatrix} 443{,}06 \\ 64{,}98 \\ 8{,}45 \\ 16{,}59 \\ 422{,}77 \end{pmatrix}$$

¿Cuáles son los parámetros que debemos estimar?

$$Cov(X,Y) = \frac{\sum_1^n (x_i - \bar{x})(y_i - \bar{y})}{n}$$

- Para estimar la fuerza de la relación entre dos variables, se calcula la Covarianza entre dos variables aleatorias
- Si Cov(X,Y) mayor a 0 tienen fuerte relación positiva.
- Si Cov(X,Y) menor a 0 tienen fuerte relación negativa.
- Si Cov(X,Y) igual a 0 no están relacionados.

¿Cuáles son los parámetros que debemos estimar?

Coeficiente de correlación:

El coeficiente de correlación ρ de las variables ***X*** e ***Y*** se define por:

$$\rho = Corr(X,Y) = \frac{Cov(X,Y)}{\sigma_X.\sigma_Y}$$

Las siguientes proposiciones muestran que el coeficiente de correlación soluciona el problema de la covarianza que depende de las unidades en las que se expresa la variable.

Proposición: $-1 \leq Corr(X,Y) \leq 1$ para cualquier par de variables *X* e *Y*.

Bibliografía I


Tom M. Mitchel
Machine Learning, cap 5
McGraw-Hill Science/Engineering/Math; (March 1, 1997)