Nahuelanca Maximiliano
38.786.261

Tec. Ciencia de datos e inteligencia artificial.

Probabilidad y estadística.

Actividad integradora clase N° 4.

Problemática:

Estudiar la base de datos “colesterol en niños” para identificar cuáles son las variables que se pueden relacionar y qué tipo de relación existe entre ellas. Además, determinar las rectas de ajuste que relacionan todos los pares posibles de combinaciones entre las variables que se consideran en la base de datos. Con toda esta información, se espera que elaboren un informe escrito que presente las principales conclusiones del estudio y las utilidades de los resultados encontrados.

Desarrollo:

Con la información proporcionada podemos definiendo que:

- **Población**: Niños que padecen de colesterol.
- **Muestra**: Niños que padecen de colesterol entre 5 y 11 años de edad.
- **Unidad de observación**: Cada niño uno de los niños.
- **Variables**:
    - **Peso**: Cuantitativa continua
    - **Talla**: Cuantitativa continua
    - **Sexo**: Cualitativa no ordenable
    - **Edad**: Cuantitativa continua
    - **IMC**: Cuantitativa continua
    - **Colesterol**: Cuantitativa continua

Una vez planteados estos conceptos, empezaremos a examinar la base de datos “colesterol en niños”.

Inicialmente veremos cómo se relacionan las variables planteadas:

Cuadro 1. Análisis de correlaciones de las variables.

Variable(1)	Variable(2)	n	Pearson	p-valor
PESO	TALLA	382	0,823	<0,0001
PESO	EDAD	382	0,659	<0,0001
PESO	IMC	382	0,875	<0,0001
PESO	COLEST	382	-0,029	0,5747
TALLA	EDAD	382	0,850	<0,0001
TALLA	IMC	382	0,465	<0,0001
TALLA	COLEST	382	-0,067	0,1933
EDAD	IMC	382	0,335	<0,0001
EDAD	COLEST	382	0,003	0,9552
IMC	COLEST	382	0,004	0,9401

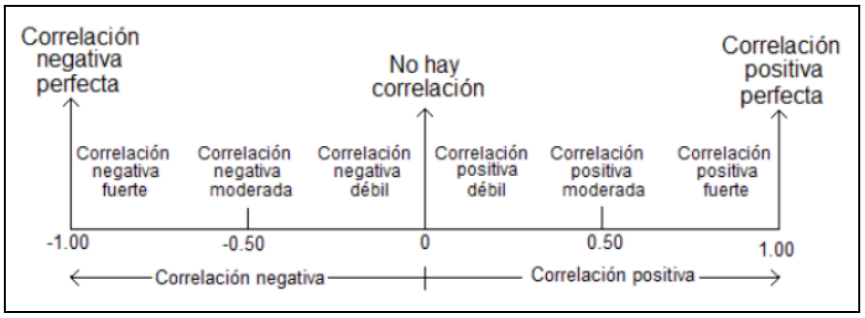
Fuente: Elaboración propia.

De acuerdo a la observación del coeficiente de correlación de Pearson, podemos tomar como estudio los siguientes pares de variables:

- Peso y talla, r = 0.8
- Peso y edad, r = 0.6
- Peso e IMC, r = 0.8
- Talla y edad. r = 0.8

Es importante mencionar que, se han tomado estos pares de variables ya que su coeficientes de Pearson es mayor a 0.5, lo que indica una correlación lineal positiva fuerte.

Figura 1: Intensidad y la dirección del coeficiente de correlación.



Fuente: Politécnico "Malvinas Argentinas"

A su vez, la variable "colesterol" no tiene una relación lineal con las demás variables estudiadas ya que sus valores están muy cerca del 0 o en negativo.

También, podemos decir que el Índice de Masa Corporal (IMC), para su cálculo toma dos valores que en este estudio tenemos como variables, a saber:

- Peso
- Talla

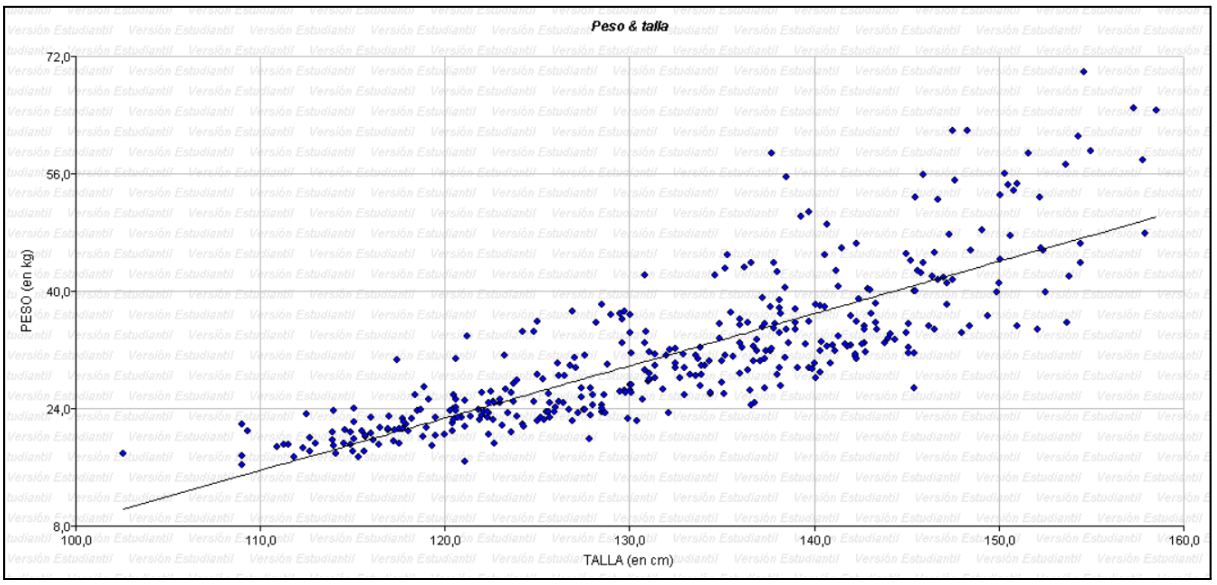
Por ello, es que se tiene una correlación fuerte de 0.8 entre Peso e IMC y una moderada de 0.4 entre Talla e IMC.

Dicho esto, seguiremos esta presentación con estos pares:

1. Peso y talla, r = 0.8
2. Peso y edad, r = 0.6
3. Talla y edad. r = 0.8

1) **Relación entre las variables PESO y TALLA**: Consideramos que la talla es la variable independiente o regresora y el peso la variable dependiente. Por ende, el peso depende de la talla.

Figura 2. Diagrama de dispersión peso y talla



Fuente: Elaboración propia.

El gráfico demuestra que efectivamente hay una correlación positiva fuerte según el coeficiente de correlación de Pearson, junto a una regresión lineal directa con su pendiente positiva.

Ahora veamos la tabla de análisis de regresión lineal.

Cuadro 2. Análisis de regresión lineal, peso y talla.

Análisis de regresión lineal

Variable	N	R^2	R^2 Aj	ECMP	AIC	BIC
PESO	382	0,68	0,68	32,82	2416,95	2428,78

Coeficientes de regresión y estadísticos asociados

Coef	Est.	E.E.	LI(95%)	LS(95%)	T	p-valor	CpMallows	VIF
const	-62,63	3,34	-69,19	-56,07	-18,76	<0,0001		
TALLA	0,71	0,03	0,66	0,76	28,29	<0,0001	800,29	1,00

Fuente: Elaboración propia.

El coeficiente muestral de determinación (R^2), indica que **el 68% de la variabilidad del peso se explica a partir de la variabilidad de la talla**. Esto concuerda con el coeficiente de Pearson el cual indicaba una correlación fuerte, pero también, existe un 32% de variabilidad del peso que se explica a partir de otros factores.

Con estos datos, podemos plantear la ecuación del modelo matemático de ajuste, donde, en este caso, Y es el peso y X la talla:

$$y = a * x + b \; | \; | \; | \;\; y = 0{,}71 * x + (-62{,}63)$$

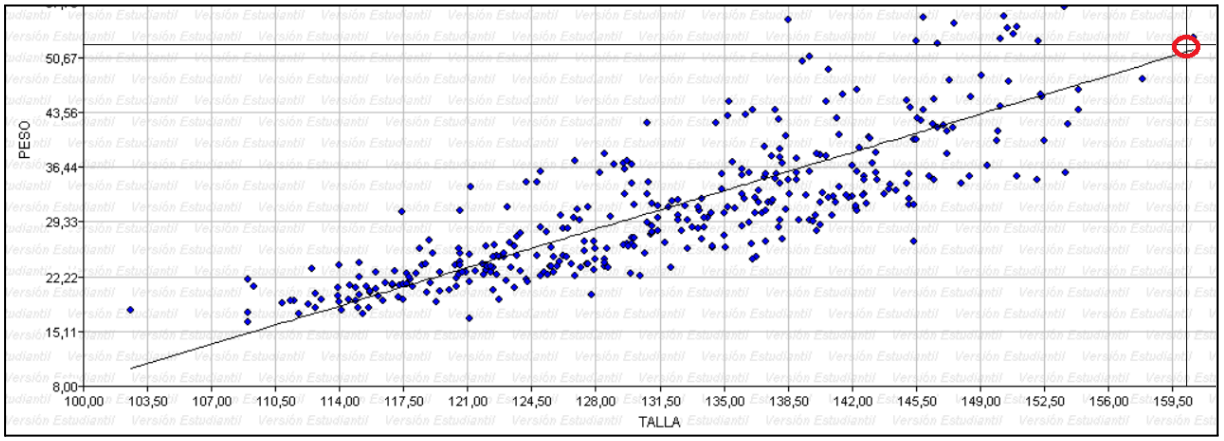
Entonces, por ejemplo, podríamos preguntarnos **¿Cuál sería el peso aproximado para un niño/a que mide 160cm?**

$$y = 0{,}71 * \mathbf{160} + (-62{,}63) \; | \; \mathbf{y = 50{,}97\ kg}$$

En el gráfico podríamos trazar las líneas y llegar al punto de intersección:

- Y (peso): 50,97
- X (talla): 160

Figura 3. Diagrama de dispersión peso y talla con predicción.



Fuente: Elaboración propia.

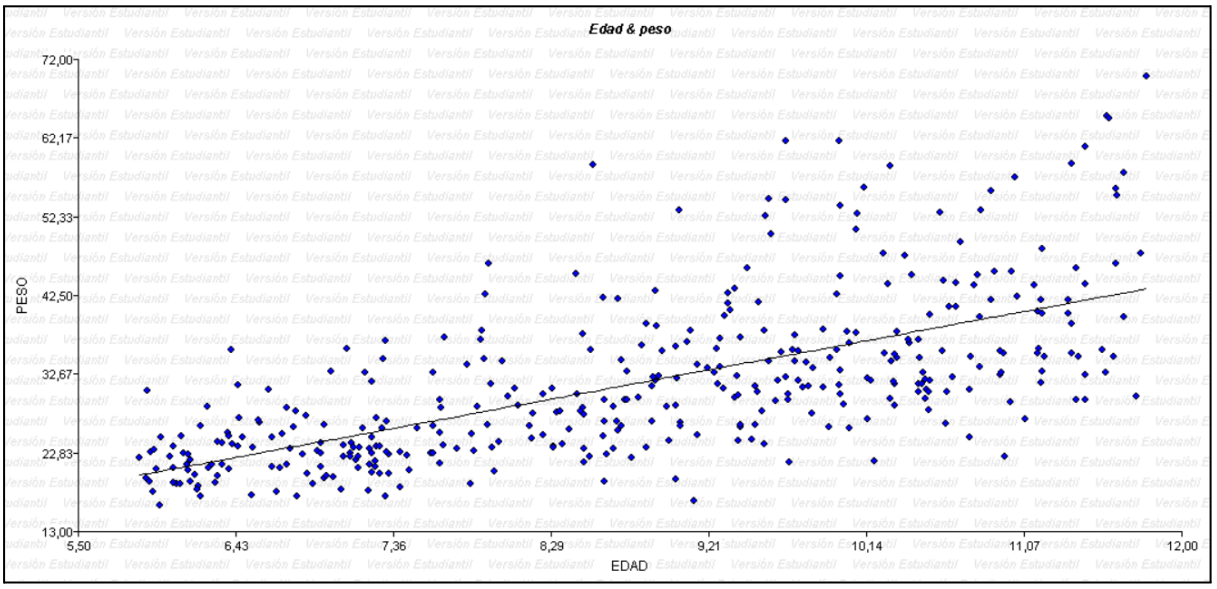
---

2) **Relación entre las variables PESO y EDAD**: Consideramos que la edad es la variable independiente o regresora y el peso la variable dependiente. Por ende, el peso depende de la edad.

   Si vemos el “Cuadro 1”, podemos observar que la correlación entre estas dos variables es de 0.6, lo que indica una correlación positiva moderada.

   Ahora graficamos el diagrama de dispersión junto a su regresión lineal.

Figura 4. Diagrama de dispersión, edad y peso.



Fuente: Elaboración propia.

El gráfico demuestra que existe una correlación positiva moderada según el coeficiente de correlación de Pearson, junto a una regresión lineal directa con su pendiente positiva.

Por ende, si analizamos la tabla de análisis de regresión lineal;

Cuadro 3. Análisis de regresión lineal, edad y peso.

Análisis de regresión lineal

Variable	N	R^2	R^2 Aj	ECMP	AIC	BIC
PESO	382	0,43	0,43	57,63	2632,60	2644,43

Coeficientes de regresión y estadísticos asociados

Coef	Est.	E.E.	LI(95%)	LS(95%)	T	p-valor	CpMallows	VIF
const	-2,99	2,05	-7,03	1,05	-1,45	0,1467		
EDAD	3,93	0,23	3,47	4,38	17,06	<0,0001	291,15	1,00

Fuente: Elaboración propia.

Efectivamente demuestra que su correlación es moderada a débil, ya que el **43% de la variabilidad del peso se explica a través de la variabilidad de la edad**. Y un 53% se explica a través de otras variables.

Ahora veremos el modelo matemático de ajuste, donde, en este caso, Y es el peso y X la edad:

$$y = a * x + b \;|\;|\;|\; y = 3.93 * x + (-2.99)$$

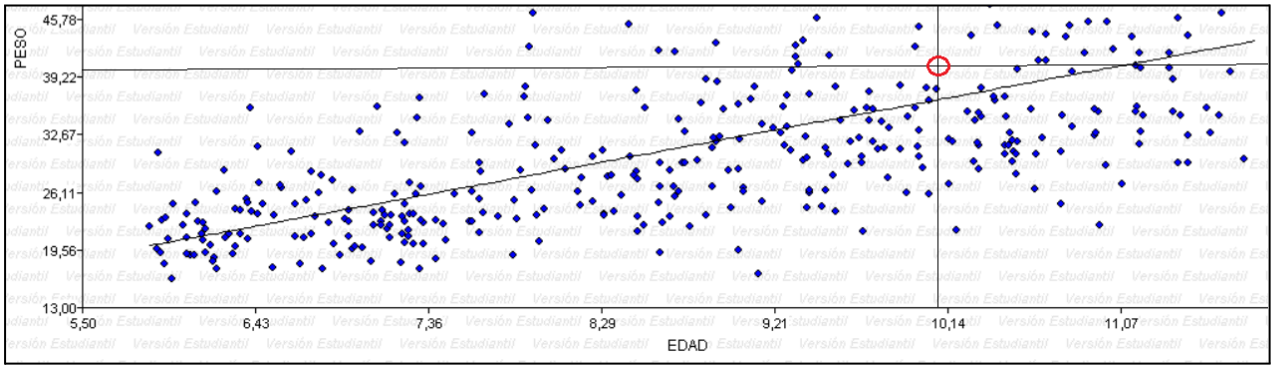
Reemplacemos los valores para probar el modelo. Podríamos preguntarnos **¿Cuál sería el peso aproximado para un niño/a que tiene 10 años de edad?**

y = 3.93 * **10** + (-2.99) | **Y = 39,3 KG**

Si trazamos las líneas y llegamos al punto de intersección:

- Y (peso): 39,3
- X (edad): 10

Figura 5. Diagrama de dispersión, edad y peso con predicción.



Fuente: Elaboración propia.

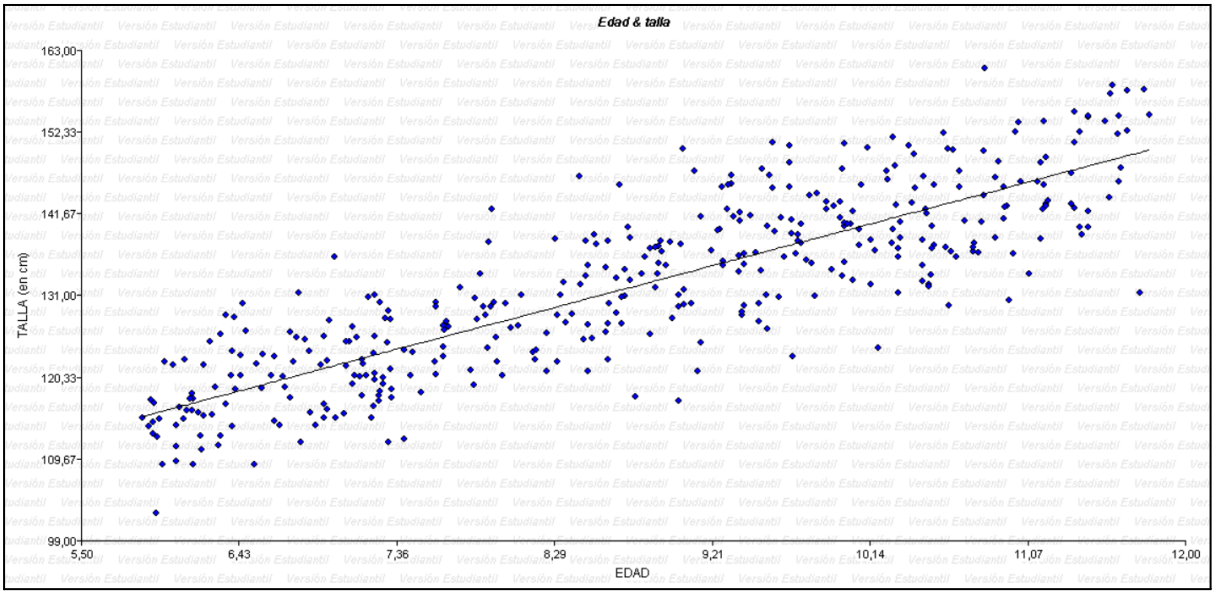
---

3) **Relación entre las variables TALLA y EDAD:** En este último caso de estudio, tomamos a la edad como variable independiente y la talla como dependiente. Sería que, la talla depende de la edad.

   Como vemos en el “Cuadro 1”, el valor de la correlación de Pearson entre estas dos variables es de 0.8, dando a entender que tienen una correlación positiva y fuerte.

   Veamos al momento de gráficas su dispersión.

Figura 6. Diagrama de dispersión, edad y talla.



Fuente: Elaboración propia.

Efectivamente, el gráfico demuestra que hay una correlación positiva fuerte según el coeficiente de correlación de Pearson, junto a una regresión lineal directa con su pendiente positiva.

Ahora veamos la tabla de análisis de regresión lineal.

Cuadro 4. Análisis de regresión lineal, edad y talla.

Análisis de regresión lineal

Variable	N	R^2	R^2 Aj	ECMP	AIC	BIC
TALLA	382	0,72	0,72	37,90	2472,72	2484,55

Coeficientes de regresión y estadísticos asociados

Coef	Est.	E.E.	LI(95%)	LS(95%)	T	p-valor	CpMallows	VIF
const	80,85	1,67	77,57	84,12	48,51	<0,0001		
EDAD	5,87	0,19	5,50	6,23	31,42	<0,0001	987,48	1,00

Fuente: Elaboración propia.

De forma afirmativa, el coeficiente muestral de determinación (R^2), indica que el **72% de la variabilidad de la talla se explica a partir de la variabilidad de la edad.** Esto concuerda con el coeficiente de Pearson el cual indicaba una correlación fuerte, de igual manera, existe un 28% de variabilidad de la talla que se explica a partir de otras variables.

Entonces, si vemos el modelo matemático de ajuste, donde Y es la talla y X la edad:

$$y = a * x + b \;|\;|\;|\; y = 5{,}87 * x + 80{,}85$$

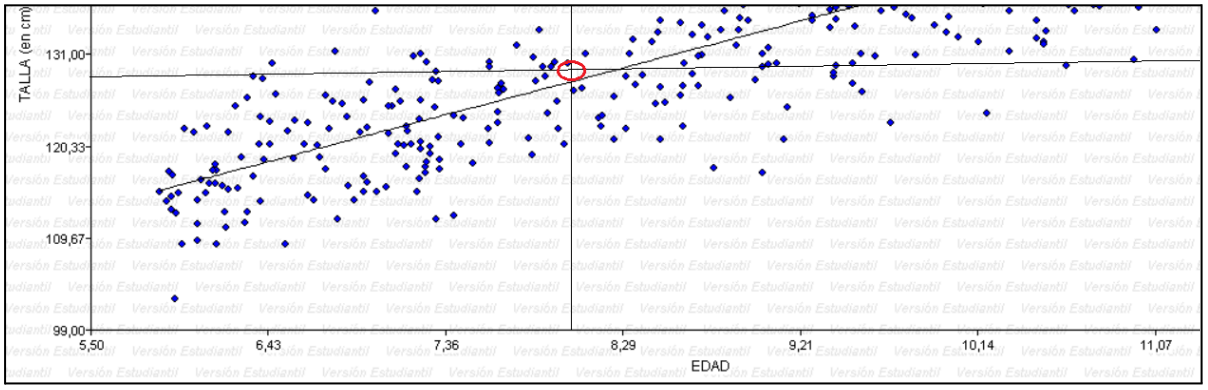
Podríamos preguntarnos **¿Cuál sería la talla (estatura) aproximada para un niño/a que tiene 8 años de edad?**

$$y = 5{,}87 * \mathbf{8} + 80{,}85 \;|\; \mathbf{Y = 127{,}81\ cm}$$

Si trazamos las líneas y llegamos al punto de intersección:

- Y (talla): 127,81
- X (edad): 8

Figura 7. Diagrama de dispersión, edad y talla con predicción.



Fuente: Elaboración propia.

Consideraciones

De los tres pares analizados en el presente informe podemos decir que el par "**Talla / Edad**" es el que presenta mayor correlación con un coeficiente muestral de determinación de 72% que avala dicha correlación. Sumado al 0.8 de correlación de Pearson y una pendiente positiva y directa en su regresión lineal.

Asimismo, es importante mencionar que, con los datos que se nos presentaron no es posible dar una respuesta concreta sobre la variable "colesterol". Pero si es posible realizar otro tipo de análisis con respecto, por ejemplo, a:

- **IMC**: Clasificar por categorías según el índice y con ello, realizar participaciones en un posterior análisis en InfoStat, a saber:
    - Peso insuficiente: IMC por debajo de 18,5
    - Normopeso: IMC entre 18,5 y 24,9
    - Sobrepeso: IMC entre 25 y 29,9
    - Obesidad tipo I: IMC entre 30 y 34,9
    - Obesidad tipo II: IMC entre 35 y 39,9
    - Obesidad tipo III (mórbida): IMC entre 40 y 49,9
    - Obesidad tipo IV (extrema): IMC mayor de 50

- **Colesterol**: Aquí también podríamos segmentar los niveles de colesterol en:
    - Entre 130 y 159 mg/dl - Límite superior del rango normal
    - Entre 160 y 189 mg/dl - Alto
    - 190 mg/dl o más - Muy alto

- **Sexo**: Es posible particionar esta variable en masculino y femenino. Partiendo de aquí, podemos realizar otros cruces de información y preguntarnos, por ejemplo:
    - ¿Cómo son los niveles de colesterol en hombres y mujeres?
    - ¿Existen diferencias sustanciales?

- **IMC/Colesterol**: También, con los segmentos definidos, podremos cruzar estas 2 variables y averiguar si existe alguna relación entre un mayor nivel de IMC y el nivel de colesterol, por ejemplo.