Matemática 4 - 2024
Trabajo Entregable:
Regresión Lineal

Integrantes:
Ajenjo, Tobías - Leg: 16.286/5
Lettieri, Juan de Dios - Leg: 16.398/2

Aclaración:Para no explicar el código, el documento será de las fórmulas utilizadas para obtener dichos resultados (algunos ejercicios si se muestra los valores) y justificando porque se usaron dichas fórmulas. Y otros archivos con los códigos.

[Link del repositorio de GitHub]

Parte 1: Predicción del valor de mercado

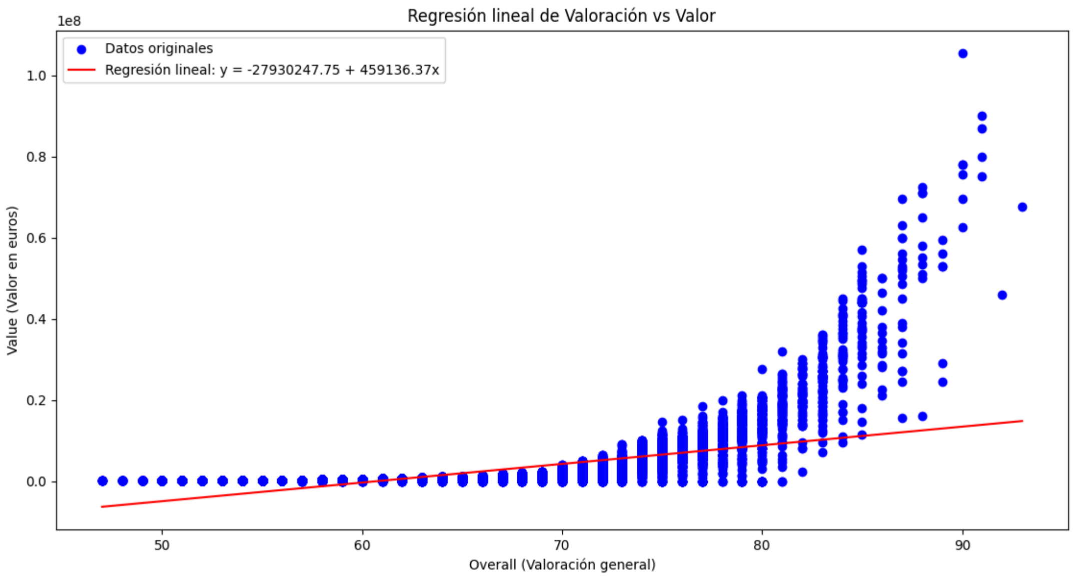
a) Recta de regresión para predecir el valor de mercado de un jugador a partir de la característica más relevante (a la que se destinará mayor proporción del presupuesto), respaldada por:

i) Prueba de significancia de regresión, coeficiente de determinación (R^2) y correlación lineal (r).

A partir del dataset brindado (*players_21.csv*) y de la consigna, sabemos que la variable **y** será **value_eur**, que representa el valor del mercado de cada jugador en euros. Para la elección de la característica más relevante, y observando que se dispone de un total de 79 características en el dataset, resulta lógico usar **overall** ya que representa un valor medio estandarizado de cada jugador.

Entonces:
Como **X** la variable Overall
Como **Y** la variable Value_eur

El trabajo será mostrar la obtención con las fórmulas utilizadas para sacar los valores.



Con los siguientes cálculos podremos realizar los cálculos como la **Prueba de regresión**, **Coeficiente de Determinación** y la **Correlación lineal**

<u>Tabla de valores obtenidos y su formula de obtencion</u>

Estadístico	Fórmula	Valor
n	Tamaño de la muestra	**18944**
$\widehat{B}_1$	$\widehat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$	**459136.3739357086**
$\widehat{B}_0$	$\widehat{\beta}_0 = \overline{y} - \widehat{\beta}_1 \cdot \overline{x}$	**-27930247.753948938**
Syy	Desviacion yi de su media $S_{yy} = \sum_{i=1}^{n}(y_i \cdot y_i) - \frac{\left(\sum_{i=1}^{n} y_i\right) \cdot \left(\sum_{i=1}^{n} y_i\right)}{n}$	**4.931878872786131e+17**
Sxx	Desviacion xi de su media $S_{xx} = \sum_{i=1}^{n}(x_i \cdot x_i) - \frac{\left(\sum_{i=1}^{n} x_i\right) \cdot \left(\sum_{i=1}^{n} x_i\right)}{n}$	**928811.2128378379**
Sxy	Covarianza $S_{xy} = \sum_{i=1}^{n}(x_i \cdot y_i) - \frac{\left(\sum_{i=1}^{n} x_i\right) \cdot \left(\sum_{i=1}^{n} y_i\right)}{n}$	**426451012333.19257**
$\overline{x}$	Media de la variable x	**65.67778716216216**
$\overline{y}$	Media de la variable y	**2224813.2918074327**
SS_r	Suma de los residuos $SS_R = S_{yy} - \widehat{\beta}_1 \cdot S_{xy}$	**2.9738871581473894e+17**
R^2	Coeficiente de determinación $R^2 = 1 - \frac{SS_R}{S_{yy}}$	**0.39700725933129566**
r	Correlacion Lineal	**0.6300851207029854**

	$r = \dfrac{S_{xy}}{\sqrt{S_{xx} \cdot S_{yy}}}$	
$\widehat{\sigma^2}$	$\widehat{\sigma^2} = \dfrac{SS_R}{n-2}$	**15699963879988.33**

Coeficiente de Determinación(R^2)= 0.39700725933129566
Correlación lineal(r) = 0.6300851207029854

Para la **prueba de regresión**,debemos ver si la pendiente b1 de la regresión es significativamente diferente de 0. Esto nos dirá si existe una relación lineal significativa entre la variable independiente X ("overall") y la variable dependiente Y (valor en euros).

Pasos del test de hipótesis

1 - Establecemos Hipótesis Nula (H0) y Hipotesis Alternativa (Ha)

H0: $\beta_1 = 0$ (no hay relación lineal)
Ha: $\beta_1 \neq 0$ (una relación lineal con pendiente distinto de 0)

2 - Nivel de significancia

1 - α = 0.95 ⇔ α = 1 - 0.95 = 0.05

3 - Elección del estadístico de prueba

T: $T = \beta_1 - \theta / \sigma^2/\sqrt{Sxx}$ ~ t $(n - 2)$ bajo H0

4- Definimos nuestra regla de decisión:

$$Regla: \begin{cases} \text{rechazar} \quad H_0 \quad si \quad |T| > t_{\frac{\alpha}{2},\, n-2} \\ \text{aceptar} \quad H_0 \quad si \quad |T| \leq t_{\frac{\alpha}{2},\, n-2} \end{cases}$$

5- Cómo tomaremos la decisión:

- Si $H_a : \beta_1 > \theta$ se rechaza $H_0 : \beta_0 = \theta$ si $T > t_{\alpha,n-2}$
- Si $H_a : \beta_1 < \theta$ se rechaza $H_0 : \beta_0 = \theta$ si $T < -t_{\alpha,n-2}$

Con nuestros valores

|T| = 111.67503525224436

Valor crítico de t: 1.9600892310950557

Como 111.67503525224436 es mayor que 1.9600892310950557

Rechazamos la hipótesis nula. Por lo tanto $\beta_1 \neq 0$ es decir la regresión es significativa

ii) Inferencias sobre los parámetros de la recta, estimando las fluctuaciones con una confianza del 95%.

Para definir un intervalo de confianza del 95%, debemos utilizar un α = 0.05

1 - α = 0.95 ⇔ α = 1 - 0.95 = 0.05

Calculamos además el valor de t necesario para definir los intervalos:

$t(\frac{\alpha}{2}, n - 2)$ = 1.9600

Valor crítico de t: 1.9600

El intervalo de confianza para β_1 estará dado entonces por:

Definición 1.8. *Un intervalo de confianza de* $100(1-\alpha)\%$ *para la pendiente* β_1 *de la línea de regresión verdadera es:*

$$\left[\hat{\beta}_1 - t_{\frac{\alpha}{2},n-2}\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}\ ;\ \hat{\beta}_1 + t_{\frac{\alpha}{2},n-2}\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}\right]$$

reemplazando nos quedaría:

$= 459136.3739357086 \; - \; 1.9600 \, / \; 18944 - 2 \;\; * \;\; \sqrt{15.699.963.879,98 \, /928811,2128378379}$
;
$= 459136.3739357086 \; + \; 1.9600 \, / \; 18944 - 2 \;\; * \;\; \sqrt{15.699.963.879,98 \, /928811,2128378379}$

El resultado final da
= [451077.74 ; 467195.01]

iii) La proporción de veces que el valor de mercado supera la incertidumbre de predicción comparada con la respuesta media del valor de mercado para una característica fija, ambas con la misma confianza y ancho mínimo.

Para poder saber cual es la proporción de veces que supera la incertidumbre de la predicción de la calidad a la de respuesta media de calidad para una característica fija, debemos saber como es el intervalo de predicción y el intervalo para la respuesta de la media.

Para el intervalo de predicción, usaremos la siguiente fórmula:

$$\varepsilon(IP) = \left(t\left(\frac{\alpha}{2}, n - 2\right) \sqrt{\widehat{\sigma}^2 \left(1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{Sxx}\right)} \right)$$

Para el intervalo para la respuesta de la media, usaremos la siguiente fórmula:

$$\varepsilon(\text{ICM}) = \left(t\left(\frac{\alpha}{2}, n - 2\right) \sqrt{\widehat{\sigma}^2 \left(\frac{1}{n} + \frac{(x^* - \bar{x})^2}{Sxx}\right)} \right)$$

Donde x^* toma el valor: $\bar{x}$.

Por lo tanto estamos utilizando un ancho mínimo, dado que el ancho del intervalo es mínimo cuando $x^* = \bar{x}$.

$$\frac{\varepsilon(IP)}{\varepsilon(ICM)} = \frac{\left(t\left(\frac{\alpha}{2}, n - 2\right) \sqrt{\widehat{\sigma}^2 \left(1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{Sxx}\right)} \right)}{\left(t\left(\frac{\alpha}{2}, n - 2\right) \sqrt{\widehat{\sigma}^2 \left(\frac{1}{n} + \frac{(x^* - \bar{x})^2}{Sxx}\right)} \right)} = \frac{\sqrt{\widehat{\sigma}^2 \left(1 + \frac{1}{n}\right)}}{\sqrt{\widehat{\sigma}^2 \frac{1}{n}}} = \frac{\sqrt{\widehat{\sigma}^2}\sqrt{1 + \frac{1}{n}}}{\sqrt{\widehat{\sigma}^2}\sqrt{\frac{1}{n}}} = \frac{\sqrt{\frac{1}{n}(n+1)}}{\sqrt{\frac{1}{n}}} =$$

$$\frac{\sqrt{\frac{1}{n}}\sqrt{(n+1)}}{\sqrt{\frac{1}{n}}} = \sqrt{(n + 1)}$$

Ahora reemplazamos el valor n y resolvemos:

$\sqrt{(18944 + 1)}$ = $\sqrt{(18945)} =$ 137,64083

Por lo tanto la proporción de veces que supera la incertidumbre de la predicción de la calidad a la de respuesta media de calidad para una característica fija utilizando un ancho mínimo es de 137,64083.

Parte 2: Ecuación de predicción del valor de mercado.

b) Ecuación para predecir el valor de mercado del jugador a partir de varias características.

Ecuación de regresión estimada para este caso, con 3 variables independientes:

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

i) Usando el método de mínimos cuadrados. Explica los indicadores obtenidos (como el coeficiente de determinación y la correlación) y proporciona una breve interpretación de los resultados.

Para la Regresión Lineal Múltiple aplicando el método de mínimos cuadrados, primero determinamos cuáles serán las variables independientes que tendremos en cuenta.

Para lograr determinar cuáles son las 3 características más fuertemente relacionadas con nuestra variable dependiente **value_eur**, se ejecutará un algoritmo que recorre todo el dataset almacenando en una lista para cada columna los valores $\boldsymbol{\beta_0}$**,** $\boldsymbol{\beta_1}$**,** $\mathbf{R^2}$ y **r**, y la ordena según el valor de **r**, de mayor a menor.

A partir de la ejecución del algoritmo se determina que las características que presentan los Coeficientes de Correlación **r** más elevados son:

- wage_eur
- overall
- potential

Estas tres características serán las que formen parte de la matriz **X**, que será de la forma:

$$\begin{array}{|c|} \hline \beta_0 \\ \hline \beta_1 \\ \hline \beta_2 \\ \hline \beta_3 \\ \hline \end{array} \qquad \begin{array}{|c|c|c|c|} \hline n & \sum x_1 & \sum x_2 & \sum x_3 \\ \hline \sum x_1 & \sum x_1^2 & \sum x_1 x_2 & \sum x_1 x_3 \\ \hline \sum x_2 & \sum x_1 x_2 & \sum x_2^2 & \sum x_2 x_3 \\ \hline \sum x_3 & \sum x_1 x_3 & \sum x_2 x_3 & \sum x_3^2 \\ \hline \end{array} \qquad \begin{array}{|c|} \hline \sum y \\ \hline \sum x_1 y \\ \hline \sum x_2 y \\ \hline \sum x_3 y \\ \hline \end{array}$$

Quedando la forma:

$$\mathbf{X\beta = Y}$$

Después de correr el algoritmo, se realizan todas las operaciones correspondientes para la obtención de los coeficientes β resultando:

Valores de la Matriz X			
18944	164355350	1244200	1346667
164355350	8743796442500	204490926470000	221331926118450
1244200	204490926470000	82645114	1675523081400
1346667	221331926118450	1675523081400	96437331

Valores de la Matriz X^{-1}			
$5.27913419 \times 10^{-5}$	$-1.60618670 \times 10^{-13}$	$-2.12115425 \times 10^{-11}$	$-1.95975662 \times 10^{-11}$
$-1.60618670 \times 10^{-13}$	$-1.85100000 \times 10^{-17}$	$2.44556000 \times 10^{-15}$	$2.25948000 \times 10^{-15}$
$-2.12115425 \times 10^{-11}$	$2.44556000 \times 10^{-15}$	$-3.23051370 \times 10^{-13}$	$2.98389740 \times 10^{-13}$
$-1.95975662 \times 10^{-11}$	$2.25948000 \times 10^{-15}$	$2.98389740 \times 10^{-13}$	$-2.75760170 \times 10^{-13}$

Valores de la Matriz Y	Valores de β
42146863000	$-1.11225541 \times 10^{6}$
6927062419767050000	1.28259427×10^{2}
52439126944600000	1.69350630×10^{4}
56757789555621000	1.56464928×10^{4}

Teniendo los valores β, podemos establecer que la ecuación de regresión será:

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

$$\hat{y} = -\ 1112255.4075055611\ +\ 128.25942667280268 \cdot x_1\ +\ 16935.06303872757\ \cdot x_2 +\ 15646.492778421403\ \cdot x_3$$

Con respecto a los indicadores obtenidos:

Coeficiente de Determinación: R^2	0.40893402 Indica que la precisión de las estimaciones serán un 40,89% eficaces con respecto al valor real esperable. El 40,89% de las variaciones del valor de los jugadores se puede explicar a través de las variables independientes elegidas.
R^2 ajustado	0.40884039 De manera similar a R^2, explica el porcentaje de variación del valor en euros de los jugadores (y) a través del conjunto de variables elegidas. Un 40,88% de la variación del valor de los jugadores se debe al

	conjunto de las variables overall, potential y wage.
Coeficiente de Correlación: r	0.63947949 Este valor señala una correlación positiva moderada entre la precisión y confiabilidad de las variables.

ii) Usando el método de descenso por gradiente. ¿Son los valores obtenidos iguales a los conseguidos mediante la resolución del sistema de ecuaciones normales? Muestra los resultados obtenidos junto con las últimas iteraciones del algoritmo. Indica los valores de los parámetros utilizados (como tasa de aprendizaje y número de iteraciones).

Utilizando el código para realizarlo con descenso de gradiente, podemos ver que los resultados **NO** son idénticos.

Aclaración: Para que funcione el código tuvimos que normalizar los elementos, asique los valores no serán los mismos mencionados anteriormente.

Con tasa de aprendizaje de 0.01

Y cantidad de iteraciones :1000

```
Resultados finales después de descenso por gradiente:
Intercepto (B0): 2224813.2880631443
Coeficiente para 'wage_eur' (B1): 3506933.6053420496
Coeficiente para 'overall' (B2): 678761.3693528511
Coeficiente para 'potential' (B3): 821406.2644608931
--------------------------------------------------
Resultados usando ecuaciones normales:
Intercepto (B0): 2224813.291807436
Coeficiente para 'wage_eur' (B1): 3507246.673070676
Coeficiente para 'overall' (B2): 678002.0799992392
Coeficiente para 'potential' (B3): 821906.4216271599
```

Podemos ver que para overall y potential son significativamente distintas.

Ahora mostrando con otros valores:

Con tasa de aprendizaje de 0.10

Y cantidad de iteraciones :2000

Iteración 1700: ECM = 6422102026281.628 Iteración 1800: ECM = 6422102026281.628 Iteración 1900: ECM = 6422102026281.628 $$ Resultados finales después de descenso por gradiente: Intercepto (B0): 2224813.2918074327 Coeficiente para 'wage_eur' (B1): 3507246.6730706766 Coeficiente para 'overall' (B2): 678002.0799992433 Coeficiente para 'potential' (B3): 821906.4216271553 -- Resultados usando ecuaciones normales: Intercepto (B0): 2224813.291807436 Coeficiente para 'wage_eur' (B1): 3507246.673070676 Coeficiente para 'overall' (B2): 678002.0799992392 Coeficiente para 'potential' (B3): 821906.4216271599	Para ver acá las últimas iteraciones con su error cuadrático medio(mostrando cada 100 elementos una). Con la cantidad de iteraciones en 2000 y con una tasa de aprendizaje del 0.10 los resultados fueron muchísimo más exactos que los mostrados con los otros valores. Esto se nota más para las variables potential y overall.

iii) Da una interpretación del criterio de corte utilizado en el algoritmo del gradiente. Explica si presenta alguna falla. Si no es una buena condición de corte, ¿puedes sugerir un criterio alternativo más eficaz?

En el ejercicio anterior solo se marcó una **cantidad de iteraciones** para su **corte**, aunque este criterio es fácil de implementar, surgen varias **fallas**:

1. **Posible sobreestimación o subestimación del número de iteraciones necesarias:**
    - Si el número de iteraciones es demasiado bajo, el algoritmo puede no haber convergido completamente, resultando en coeficientes subóptimos.
    - Si el número de iteraciones es demasiado alto, el algoritmo puede seguir iterando innecesariamente, desperdiciando tiempo y recursos computacionales.
2. **No mide la convergencia real:**
    - El número de iteraciones no refleja si el algoritmo ha llegado a una solución óptima. Es posible que después de un cierto número de iteraciones, los coeficientes cambien muy poco o nada, lo que indicaría que el algoritmo ya ha convergido, pero el criterio basado en iteraciones seguiría ejecutándose.

Una **sugerencia** de un criterio alternativo más eficaz podría ser detener el algoritmo cuando el **gradiente o el cambio en la función de costo** (error cuadrático medio) sea muy chico. Este enfoque asegura que el algoritmo se detiene solo cuando ha convergido a una solución óptima o cercana.

Hicimos un algoritmo que cuando el cambio absoluto en el error cuadrático medio (MSE) entre iteraciones sucesivas sea menor que un umbral muy pequeño, por ejemplo, `1e-6`.

Esto indicaría que el modelo ha dejado de mejorar significativamente y ha alcanzado la convergencia.

```
Iteración 200: ECM = 6422102026281.768
El algoritmo ha convergido en la iteración 218 con ECM = 6422102026281.639
$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$
Resultados finales después de descenso por gradiente:
Intercepto (B0): 2224813.2918074327
Coeficiente para 'wage_eur' (B1): 3507246.6255244124
Coeficiente para 'overall' (B2): 678002.2102654438
Coeficiente para 'potential' (B3): 821906.3299061068
--------------------------------------------------
Resultados usando ecuaciones normales:
Intercepto (B0): 2224813.291807436
Coeficiente para 'wage_eur' (B1): 3507246.673070676
Coeficiente para 'overall' (B2): 678002.0799992392
Coeficiente para 'potential' (B3): 821906.4216271599
```

Los valores no son idénticos pero son muchos más aproximados con un total de 218 iteraciones con una tasa de aprendizaje del 0.10.
Cuando antes con iteraciones hicimos 1.782 demás.

Parte 3: Comportamiento del método de descenso por gradiente

c) Convergencia del método de descenso por gradiente.
Explicar si el método siempre converge al mínimo de la función. En caso contrario, proporciona un contraejemplo para ilustrar este comportamiento.

Este método **no siempre** garantiza la convergencia al mínimo global de la función.

Dicha situación se puede presentar por varios motivos:

- **Selección del Punto Inicial**:
  Cuando empezamos el algoritmo de optimización, el punto donde empezamos puede influir en el resultado.
  Si elegimos un punto inicial que se encuentra bastante alejado del punto mínimo global, puede ocurrir que el algoritmo converja, se detenga en un mínimo local y no global.

- **Selección valor de tolerancia o tasa de aprendizaje**
  La tasa de aprendizaje determina el tamaño de los pasos que se dan al actualizar los valores durante el proceso de optimización.

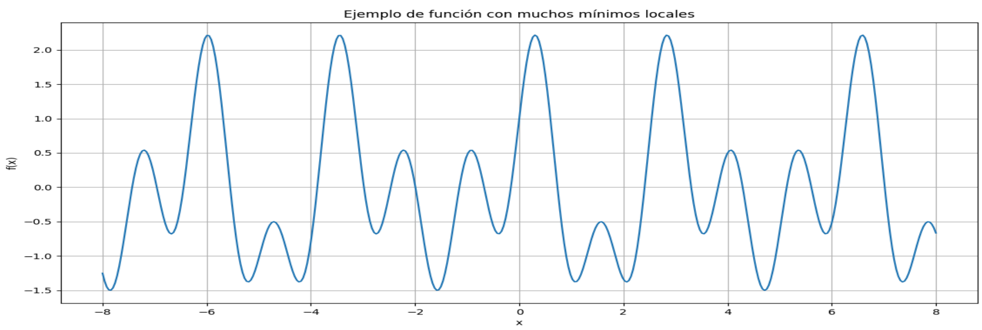
- Si la tasa de aprendizaje es excesivamente alta, el algoritmo puede dar saltos demasiado grandes, lo que puede llevar a que se aleje del mínimo global en lugar de acercarse a él.
- Si la tasa de aprendizaje es excesivamente baja, el algoritmo puede dar saltos tan chicos que puede quedar atrapado en un mínimo local sin llegar al mínimo global.

Por lo tanto, el **punto de inicio** y la elección de la **tasa de aprendizaje** son factores críticos que afectan la **convergencia**.

Contraejemplo:

Para funciones no convexas, el método puede no converger al mínimo global. Puede converger a un mínimo local, lo que significa que el algoritmo puede encontrar un punto donde la función no disminuye más, pero que no es el mínimo absoluto de la función.

- Considerando una función con una combinación de funciones sinusoidales y polinómicas. Esta función tiene múltiples mínimos locales y un mínimo global. Dependiendo del punto de inicio y la tasa de aprendizaje, el método de descenso por gradiente podría converger a un mínimo local en lugar del mínimo global.



El método de descenso por gradiente no siempre converge al mínimo global de una función no convexa

.

Los resultados dependen del punto de inicio, la tasa de aprendizaje y la forma de la función.

En este caso la función tiene múltiples mínimos locales. Dependiendo del punto de inicio, el algoritmo puede converger a uno de estos mínimos sin alcanzar el mínimo global.

Un contraejemplo pero con una función puede ser:f(x)=$(x - 2)^2$

Aca por ejemplo la tasa de aprendizaje es tan alta que jamas convergerá en el minimo global.