JUEGOS OLÍMPICOS

Oscar Angarita - James Romero - Andres Lindarte - Santiago Sandoval

PREGUNTAS:

- ¿Cómo ha sido el desempeño de colombia en los juegos olímpicos de verano desde su primera participación?
- 2. ¿Qué deporte se debería incentivar más en el país para poder conseguir mejores resultados en los juegos olímpicos?
- 3. ¿Cómo ha sido el rendimiento de la delegación de weightlifthing (levantamiento de pesas) colombiana a comparación de rusia?
- 4. ¿Dado N(cantidad determinada) atletas que compitan en representación de un NOC, cuantos de ellos podrían obtener una medalla de cualquier tipo ?

1. Comprensión del negocio

1.1 Objetivo de negocio

Conocer el trayecto histórico de Colombia frente a los juegos olímpicos y cuales son sus principales fortalezas.

1.2 Evaluación de la situación

Dataset con 271 Mil observaciones y 15 variables sobre participaciones en los juegos olímpicos de los últimos 120 años

1.3 Determinar objetivos de la minería de datos

Comparar las variables de desempeño individual a través de los años, eventos y resultados en relación con Colombia y el mundo

1.4 Producir plan de trabajo

Recolección, clasificación y comprensión de datos Ordenar las observacione según el desempeño en cantidad de medallas Comparar rendimiento entre paises Analizar la diferencia entre medallistas y no medallistas Predecir cantidad de medallas

2. Comprensión de los datos

2.1 Recolectar datos iniciales

Dataset: "120 años de historia olímpica atletas y resultados", publicado por el científico Randi H Griffin, que obtuvo la información sin formato de la página (<u>www.sports-reference.com</u>).

via: <u>www.kaggle.com</u>

2.2 Describir los datos

Documento csv 271116 filas x 15 columnas

ID - Número único por atleta Name - Nombre y apellido Sex - M(Male) o F(Female) Age - Edad del participante Height - Altura en centímetros Weight - Peso en Kilogramos Team - Nombre del equipo/país NOC - Nomenclatura oficial otorgada por el comité olímpico a cada país en un código de tres letras

Games - Año y temporada del evento

Year - año numero entero
Season - temporada dividido entre
(Summer) / (winter)
City - Ciudad de nacimiento del
atleta
Sport - Deporte del atleta
Event - Eventos por deporte
Medal - Oro/Plata/Bronce y sin
medalla

2.3 Explorar los datos

(206165, 15)

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 206165 entries, 0 to 271115
Data columns (total 15 columns):
         206165 non-null int64
ID
         206165 non-null object
Name
         206165 non-null object
Sex
         206165 non-null float64
Age
Height
         206165 non-null float64
         206165 non-null float64
Weight
Team
         206165 non-null object
NOC
         206165 non-null object
         206165 non-null object
Games
Year
         206165 non-null int64
         206165 non-null object
Season
         206165 non-null object
City
         206165 non-null object
Sport
         206165 non-null object
Event
         30181 non-null object
Medal
dtypes: float64(3), int64(2), object(10)
memory usage: 25.2+ MB
```

Head() Original

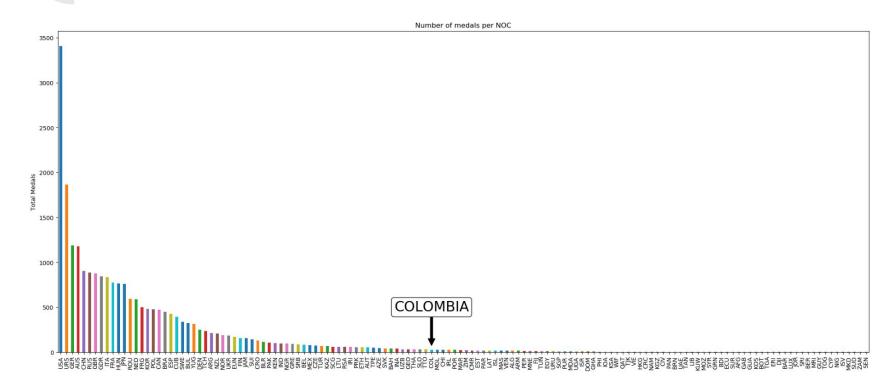
	ID	Name	Sex	Age	Height	Weight	Team	NOC	Games	Year	Season	City	Sport	Event	Medal
0	1	A Dijiang	M	24.0	180.0	80.0	China	CHN	1992 Summer	1992	Summer	Barcelona	Basketball	Basketball Men's Basketball	NaN
1	2	A Lamusi	M	23.0	170.0	60.0	China	CHN	2012 Summer	2012	Summer	London	Judo	Judo Men's Extra-Lightweight	NaN
4	5	Christine Jacoba Aaftink	F	21.0	185.0	82.0	Netherlands	NED	1988 Winter	1988	Winter	Calgary	Speed Skating	Speed Skating Women's 500 metres	NaN
5	5	Christine Jacoba Aaftink	F	21.0	185.0	82.0	Netherlands	NED	1988 Winter	1988	Winter	Calgary	Speed Skating	Speed Skating Women's 1,000 metres	NaN
6	5	Christine Jacoba Aaftink	F	25.0	185.0	82.0	Netherlands	NED	1992 Winter	1992	Winter	Albertville	Speed Skating	Speed Skating Women's 500 metres	NaN

Describe() Original

	ID	Age	Height	Weight	Year
count	206165.000000	206165.000000	206165.000000	206165.000000	206165.000000
mean	68616.017675	25.055509	175.371950	70.688337	1989.674678
std	38996.514355	5.483096	10.546088	14.340338	20.130865
min	1.000000	11.000000	127.000000	25.000000	1896.000000
25%	35194.000000	21.000000	168.000000	60.000000	1976.000000
50%	68629.000000	24.000000	175.000000	70.000000	1992.000000
75%	102313.000000	28.000000	183.000000	79.000000	2006.000000
max	135571.000000	71.000000	226.000000	214.000000	2016.000000

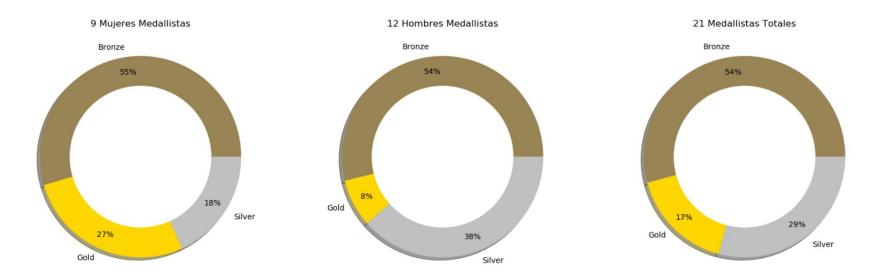
2.3 Explorar los datos

Cantidad de medallas en Colombia vs resto del mundo



2.3 Explorar los datos

Cantidad de medallistas en Colombia por sexo



2.4 Verificar la calidad de los datos

Se corroboró la veracidad de los datos al comparar la fuente principal (www.sports-reference.com) con documentación web sobre la cantidad de medallas, países participantes y datos atípicos que parecen errores y no datos verídicos, por ejemplo, se encuentran participantes jóvenes con una edad mínima de 11 años o de avanzada edad hasta los 97 años, ambos son casos reales de atletas en gimnasia y equitación respectivamente.

3. Preparar los datos

3.1 Seleccionar

En cuanto a variables todas se conservan, ya que cada una de estas es pertinente para diferentes análisis del dataset. "Year", "NOC", "ID", "Sport", "Medals" son las columnas que más se destacan ya que permiten hacer análisis del rendimiento de los atletas a través de los años, según las medallas, el país o el deporte; las demás variables permiten complementar el tipo de atleta que participan en los eventos olímpicos.

3.2 Limpiar

Inicialmente las observaciones se organizaron por año de forma ascendente lo que da una perspectiva de cambio a través del tiempo y también se realizó una limpieza de una minoría de atletas que les faltaba información de peso, altura o deporte, por lo cual se sacaron del dataset, ya que generaban más incongruencias que información general.

3.3 Construir

Los valores predictores son cantidad de asistentes a los juegos olímpicos contra aquellos que ganaron medalla, lo que permite visualizar una tasa de éxito en los juegos según la cantidad de asistentes en un deporte. Para esto se sacaron dos dataframes diferentes uno agrupando los "ID" únicos por año y el otro "Medallistas" agrupados por año, luego al obtener la cantidad de los dos "por año" se eligen las dos variables para generar una predicción.

3.4 Integrar

Para responder las preguntas descriptivas era necesario centrar el data frame en Colombia y posteriormente en el deporte "Weightlifting" lo que permite graficar el rendimiento de Colombia en los Olímpicos.

3.5 Dar formato

La variable de "Medals" le asignaba a los no ganadores de medallas un valor Nulo, para términos de mejor visualización se cambiaron todos aquellos nulos por "No medal"; también se renombraron algunas columnas para especificar las unidades de medida que manejan los datos, estas fueron "Age", "Height", "Weight" y se cambiaron por "Age (Years)", "Height (cm)", "Weight (kg)". Todos los demás valores string del dataset tienen el formato adecuado.

4. Modelar

4.1 Seleccionar modelos

El modelo de ciencia de datos escogido es Regresión lineal ya que el problema de negocio es de carácter predictivo. La regresión lineal es una técnica de análisis predictivo básica que utiliza datos históricos para predecir una variable de salida.

4.2 Diseñar pruebas

-La variable de entrada o predictor es la variable (s) que ayuda a predecir el valor de -la variable de salida. Se conoce comúnmente como X .

La variable de salida es la variable que queremos predecir. Se conoce comúnmente como Y .

Para estimar Y usando regresión lineal, asumimos la ecuación:

$$Y_e = \alpha + \beta X$$

donde Y _e es el valor estimado o predicho de Y basado en nuestra ecuación lineal.

Nuestro objetivo es encontrar valores estadísticamente significativos de los parámetros α y β que minimicen la diferencia entre Y y Y $_{\rm e}$.

Si somos capaces de determinar los valores óptimos de estos dos parámetros, entonces tendremos la línea de mejor ajuste que podemos utilizar para predecir los valores de Y , dado el valor de X .

4.2 Diseñar pruebas

-La variable de entrada o predictor es la variable (s) que ayuda a predecir el valor de -la variable de salida. Se conoce comúnmente como X .

La variable de salida es la variable que queremos predecir. Se conoce comúnmente como Y .

Para estimar Y usando regresión lineal, asumimos la ecuación:

$$Y_e = \alpha + \beta X$$

donde Y _e es el valor estimado o predicho de Y basado en nuestra ecuación lineal.

Nuestro objetivo es encontrar valores estadísticamente significativos de los parámetros α y β que minimicen la diferencia entre Y y Y $_{\rm e}$.

Si somos capaces de determinar los valores óptimos de estos dos parámetros, entonces tendremos la línea de mejor ajuste que podemos utilizar para predecir los valores de Y , dado el valor de X .

4.3 Desarrolla modelos

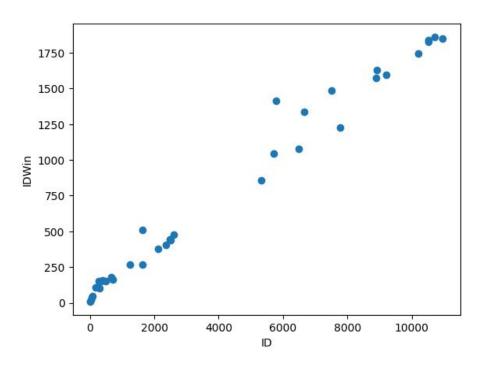
Después de realizar los dos Datasets, Cantidad de participantes por año (izquierda) y Cantidad de ganadores por año (derecha) se procede a realizar la regresión lineal entre cantidad de participantes y cantidad de

medallistas.

	Year	ID	20	Year	ID
0	1896	11	0	1896	9
1	1900	34	1	1900	20
2	1904	52	2	1904	34
3	1906	77	3	1906	45
4	1908	187	4	1908	106
5	1912	302	5	1912	103
6	1920	275	6	1920	150
7	1924	408	7	1924	158
8	1928	487	8	1928	154
9	1932	348	9	1932	151
10	1936	672	10	1936	179

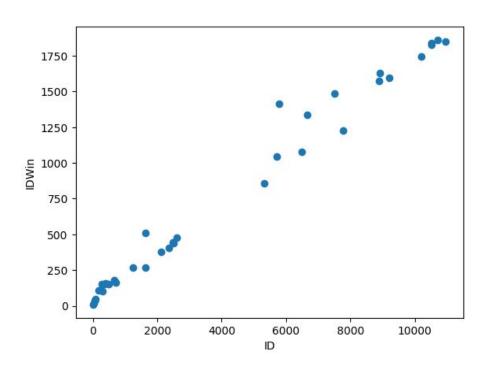
4.3 Desarrolla modelos

Gráfica de relación entre Cantidad de asistentes por año (Eje x) versus Cantidad de medallistas por año (Eje y)



4.3 Desarrolla modelos

Regresión lineal



4.4 Valorar modelos

Para una cantidad de 500 participantes:

```
lr.predict([[500]])
```

La regresión lineal predice una cantidad de medallistas de 145, con un score del 98%, es decir que el programa retorna un nivel de fiabilidad de casi el 100%

```
array([145.43792509])
```

0.9815923529980868

5.Evaluar

5.1 Evaluar resultados

Según la pregunta ¿Dado N(cantidad determinada) atletas que compitan en representación de un NOC, cuantos de ellos podrían obtener una medalla de cualquier tipo? se obtiene un resultado satisfactorio de predicción con un score bastante alto, por lo cual se podría considerar la regresión lineal un modelo exitoso de predicción para esta pregunta de negocio.

5.2 Revisar el proceso

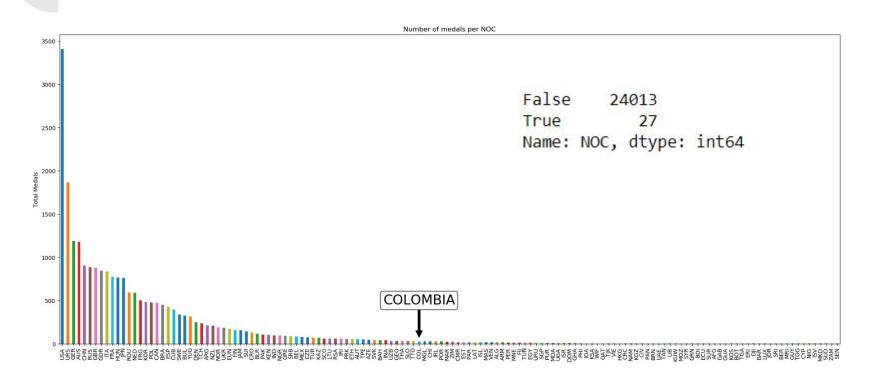
Se tuvieron en cuenta todas las variables y observaciones necesarias para realizar la predicción mediante la regresión lineal

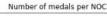
5.3 Trabajo futuro

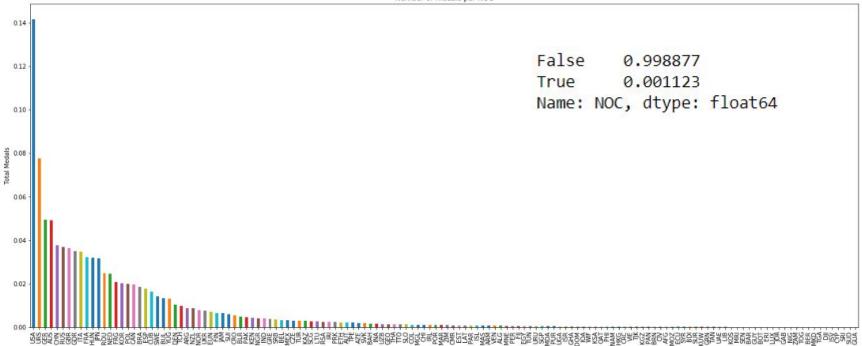
La pregunta de negocio se planteó a nivel global dado que en Colombia hay muy poca tasa de medallistas con relación a los participantes, por lo tanto actualmente no se puede obtener una predicción nacional que potencie la posibilidad de una mayor cantidad de medallas, en el futuro con mayor cantidad de datos nacionales se logrará una predicción más pertinente a la empresa interesada.

6.Conclusiones

¿Cómo ha sido el desempeño de Colombia en los juegos Olímpicos de verano desde su primera participación?





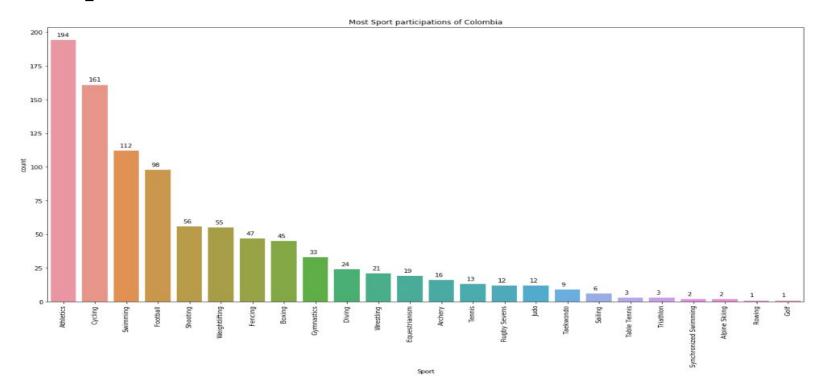


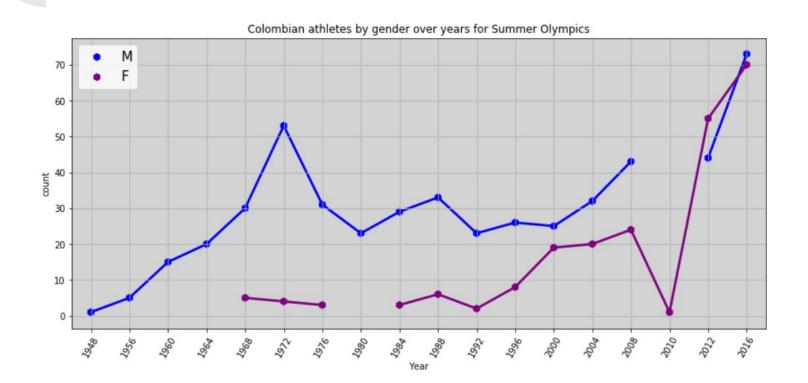
Info() del dataset de los atletas olímpicos Colombianos

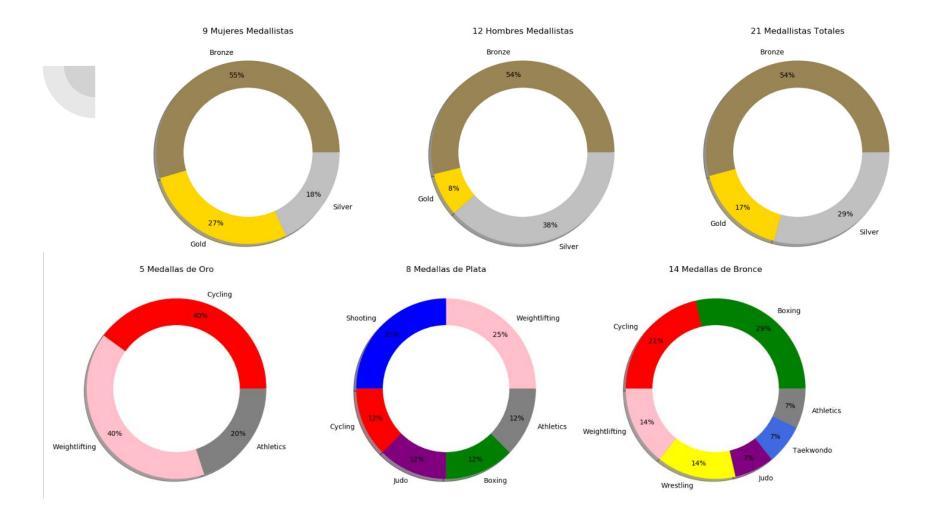
	ID	Age (Years)	Height (cm)	Weight (kg)	Year
count	945.000000	945.000000	945.000000	945.000000	945.000000
mean	63965.737566	24.985185	169.752381	64.868254	1994.205291
std	39962.490656	6.017963	9.047547	10.337798	18.654422
min	574.000000	12.000000	138.000000	38.000000	1948.000000
25%	25877.000000	21.000000	164.000000	58.000000	1976.000000
50%	74563.000000	24.000000	170.000000	64.000000	2000.000000
75%	99964.000000	28.000000	175.000000	71.000000	2012.000000
max	135441.000000	52.000000	205.000000	110.000000	2016.000000

Colombia se encuentra en la posición 65 del ranking mundial de cantidad de medallas con 27 medallas, lo que representa el 0.0011 del total global. No es un desempeño alto, ya que la participación es relativamente baja con 945 atletas desde su primera participación en el 1948 y solo ha conseguido 27 medallas, es decir solo el 2.8% de participantes son medallistas (hasta el 2016).

¿Qué deporte se debería incentivar más en el país para poder conseguir mejores resultados en los juegos olímpicos?



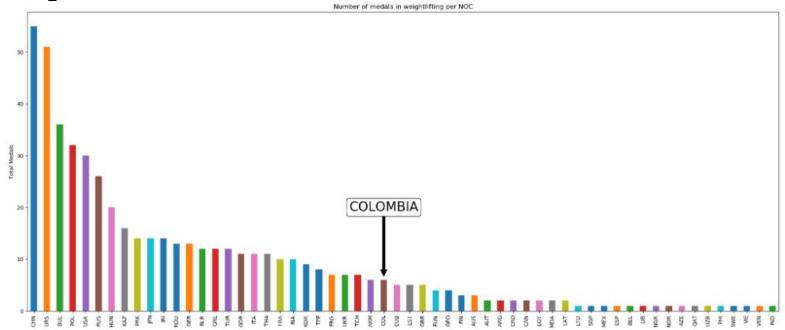




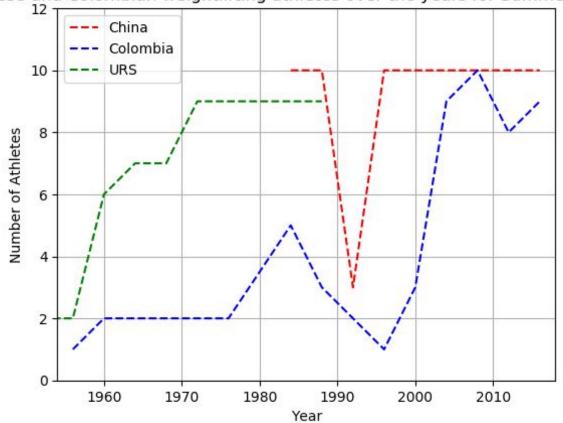
Ciclismo y Weightlifting tienen ambos 6 medallas, Weightlifting tiene 2 de plata, 2 de oro y 2 de bronce, comparado a ciclismo que tiene 1 de plata, 2 de Oro y 3 de bronce, además Weightlifting tiene casi un tercio de participantes de los que tiene ciclismo, es decir, con menor cantidad de participantes está ganando un mayor valor de medallas, lo que se traduce a una mayor tasa de éxito.

Como dato adicional, las mujeres aún con menor cantidad de participantes a través de los años ha ganado casi la misma cantidad de medallas que los hombres, pero de un mayor valor.

¿Cómo ha sido el rendimiento de la delegación de weightlifting(levantamiento de pesas) colombiana a comparación de las potencias mundiales en dicho deporte?

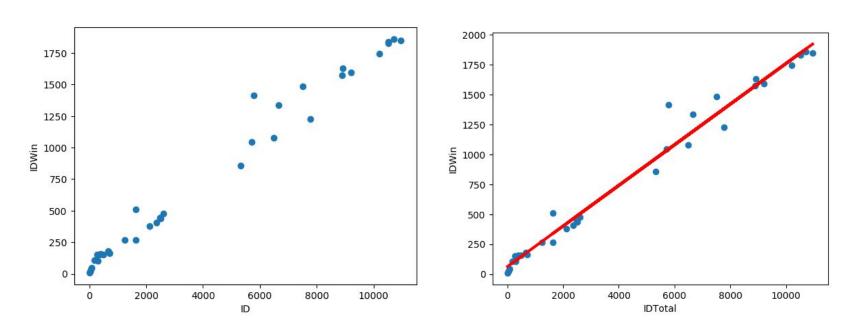


Chinese and Colombian weightlifting athletes over the years for Summer Olympics



Mundialmente Colombia no está ubicado entre los mejores puestos, se encuentra en la posición ... aun así en la gràfica de cantidad de atletas a través de los años comparado con las potencias en Levantamiento de pesas se evidencia un incremento sustancial desde el 96, este crecimiento es proporcional al rendimiento en medallas, todas siendo entregadas a Colombia desde el año 2000 en adelante. Bronce (2004), Plata (2008,2008,2012), Oro (2000, 2016). Por lo tanto podemos argumentar que el Levantamiento de pesas es un deporte con gran rendimiento en los Olímpicos para Colombia.

¿Dado N(cantidad determinada) atletas que compitan en representación de un NOC, cuantos de ellos podrían obtener una medalla de cualquier tipo?



La regresión lineal predice una cantidad de medallistas de 145, con un score del 98%, es decir que el programa retorna un nivel de fiabilidad de casi el 100%, se obtiene un resultado satisfactorio de predicción con un score bastante alto, por lo cual se podría considerar la regresión lineal un modelo exitoso de predicción para esta pregunta de negocio.

JUEGOS OLÍMPICOS

Oscar Angarita - James Romero - Andres Lindarte - Santiago Sandoval