**1.Dive Deeper**

I have used logistic regression to try to predict the medal that a ticket will win based on the variables Weight, Height and Age.

Next I will describe the process that I have carried out.

The query that was used to extract the data is the following:

Interfaz de usuario gráfica, Texto

Descripción generada automáticamente

Here you can see some important statistics from the sample:

Tabla

Descripción generada automáticamente

Before starting to process the data set, I made some visualizations that can often help to better understand the characteristics of the information we are working with and their correlation. First I visualized in history format the four input Features with names "Age", "Height", "Wight" and "Medal" we can see graphically between which values ​​their minimums and maximums are understood and in which intervals they concentrate the greatest density of records .

Gráfico, Histograma

Descripción generada automáticamente

**We create the Logistic Regression Model**

Now we load the variables of the 4 input columns in X excluding the column "Medal" with the drop () method. Instead we add the column "Medal" in the variable y. We run X.shape to check the dimension of our matrix with input data of 2000 records by 4 columns.

Texto

Descripción generada automáticamenteInterfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

Once our model is compiled, we make it classify our entire set of X inputs using the "predict (X)" method and we review some of its outputs and see that it matches the actual outputs of the data.

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

**Classification (or prediction) of new values**

It can be seen how the model predicted medal 3 (Bronze) for a completely new record. This value 3 matches the database record.

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

**2.Go Broader**

I was ignoring the NOC column that had a small error, there were RUS and URS values ​​that referred to Russia. This made my metric turn out wrong. So what I did was update all URS values ​​by RUS in both tables (athletes, regions).

So far I have discovered the following statements regarding the data sets.

• The age of most athletes in games is between 24 and 26 years old.

• The height of the athletes is 171 to 178 cm.

• Weight is 64 to 72 kg.

Also, regarding new data relationships. I believe that physical characteristics such as weight and age of an athlete influence whether or not to obtain medals.

On the other hand, another analysis that I am thinking of doing and it catches my attention is a machine learning model to predict which medals will be won the most in future events of the Olympic games.

**3.New Metric**

I have created 2 metrics: - Top 10 years with the most medals - Top 10 countries with the most medals

The reason for these metrics is that it instantly reveals some interesting data, such as: - In 2008 more medals were awarded - The country with the most medals is the USA

The graphs that were analyzed in this section are shown below.

**Metric 1: Top 10 years with the most medals**

Tabla

Descripción generada automáticamenteGráfico, Gráfico de barras

Descripción generada automáticamente

**Metric 2: Top 10 countries with the most medals**

Tabla

Descripción generada automáticamente

Gráfico, Gráfico de barras

Descripción generada automáticamente