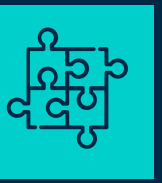
CODERHOUSE

PROYECTO DATA SCIENCE

ANÁLISIS DE LOS ATLETAS QUE COMPITEN EN LOS JUEGOS OLÍMPICOS Y SU OBTENCIÓN DE MEDALLAS.

Pablo Tomás Fernández
Pedro del Campo
Tutor: Alfredo Parente

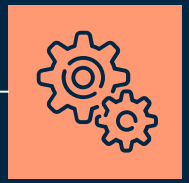
DESCRIPCIÓN DEL PROYECTO



01

PROBLEMA

Preguntas, objetivos, audiencia y fuentes.



02

PROCESO

Análisis de datos y desarrollo de modelos de ML.

03

RESULTADOS

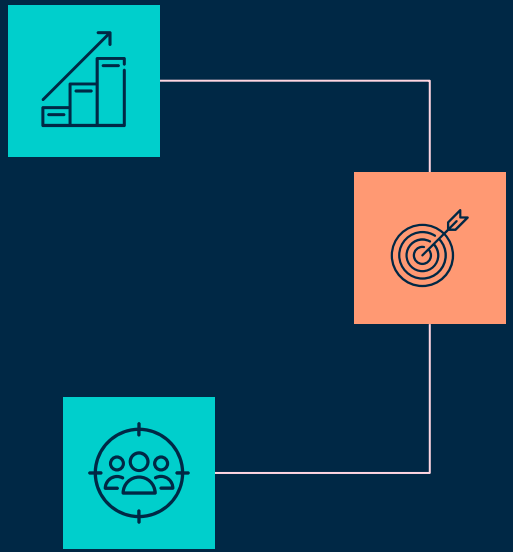
Elección del modelo y conclusiones.

PROBLEMA 01

01 Problema

TEMÁTICA

Se analizará la relación entre distintas variables de deportistas olímpicos tales como relación sexo, edad, peso, altura, país, disciplina y edición en la que compitió y qué medalla ganó.



AUDIENCIA

Creemos que esta información puede ser beneficiosa para atletas, entrenadores, nutricionistas, directores deportivos y todo aquel que esté interesado en ganar una competencia deportiva.

OBJETIVO

Crear un sistema que permita detectar cuáles son los pesos, altura y edad que tengan mayor probabilidad de obtener una medalla.

01 Problema

METADATA

Columna	Data type	Descripción del campo
athlete_id	int	id del atleta. Es uno por individuo, se puede repetir si el mismo participa en varias disciplinas o ediciones.
name	object	nombre del atleta.
sex	object	Male ('Masculino') o Female ('Femenino').
born	datetime	fecha de nacimiento.
height	float	altura (cm)
country	object	país al que representan.
country_noc	object	comité olímpico nacional al que representa. Código de tres letras
medal	object	qué medalla obtuvo ('Gold', 'Silver', 'Bronze', NaN)
isTeamSport	bool	si es un deporte de equipo o no.
event_title	object	disciplina que realiza. Ej.
sport	object	deporte que realiza. Ej.
start_date	datetime	fecha en que comenzaron esa edición de los JJOO.
Year	float	año de realización de los JJOO.
Edition	object	Si los JJOO fueron de verano o invierno.
Q_athlete_participants	float	la cantidad de atletas que compiten en esa disciplina. Campo realizado por feature engineering.
Q_country_participants	float	la cantidad de atletas que compiten representando a ese país.
Weight	float	peso (kg).
posicion	float	posición obtenida si es que la disciplina tiene posiciones finales.
athlete_years	float	edad (años). Campo realizado por feature engineering.
Medal_Bool	int	1 (obtuvo medalla) o 0 (no obtuvo medalla).
posicion_rel	float	normaliza la posición utilizando la cantidad de participantes. Obteniendo un número entre 0 y 1. Campo obtenido por feature engineering.
rango_athlete_years	object	edades pero divididas en rangos (Label encoding). Campo obtenido por feature engineering.
rango_height	object	alturas divididas en rangos (Label encoding). Campo obtenido por feature engineering.
rango_Weight	object	pesos divididos en rangos (Label encoding). Campo obtenido por feature engineering.

PREGUNTAS

¿Qué edades, peso y altura suelen tener los olimpistas?

¿Cuál es la relación entre estas variables?

¿Qué disciplina es la que tiene mayor cantidad de participantes?

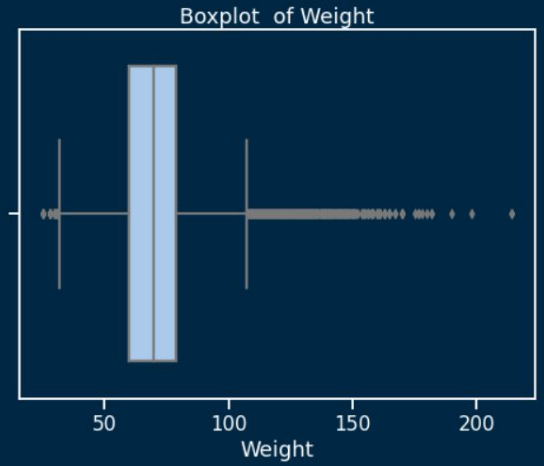
¿Cuáles son los parámetros ideales para conseguir buenos resultados?

PROCESO

02

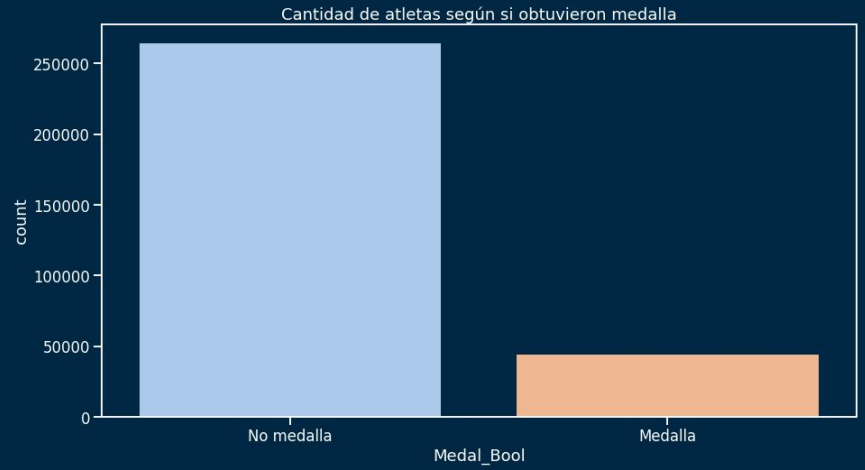
ANÁLISIS EXPLORATORIO DE DATOS (EDA)

Análisis univariado



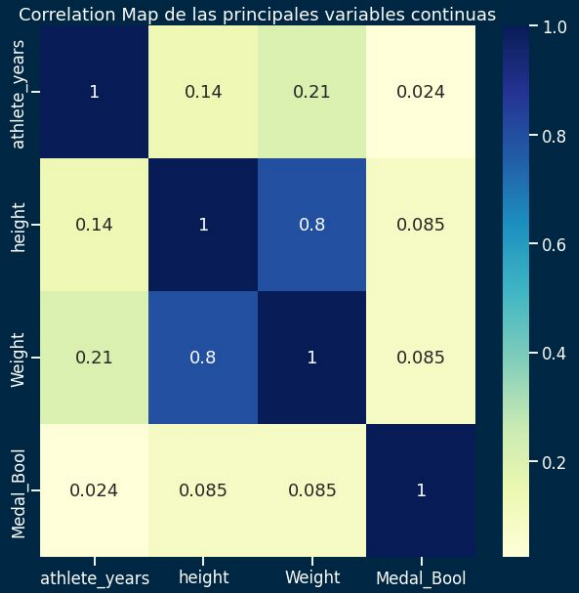
El peso, al igual que la altura y la edad, tienen una amplio rango y gran cantidad de outliers debido a las distintas disciplinas y morfologías corporales

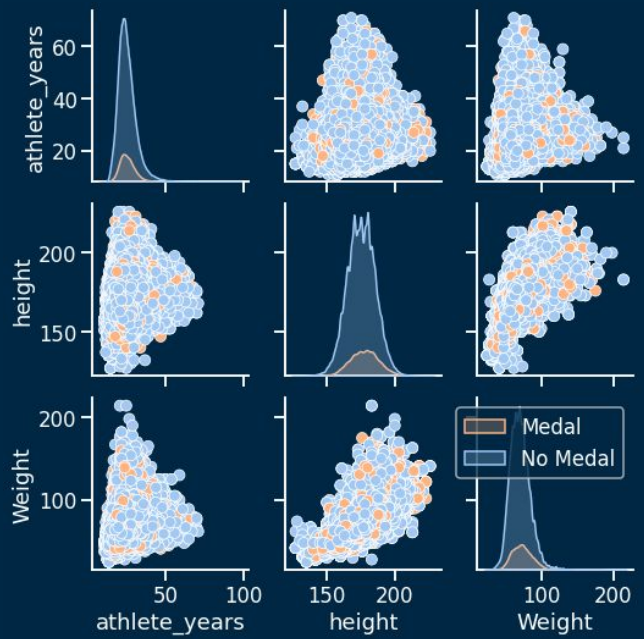
La mayoría de los atletas son de género masculino, a pesar de que en el último tiempo se han emparejado ambos grupos.
Al mismo tiempo, vemos que la mayoría de los atletas, no ganaron una medalla.



ANÁLISIS EXPLORATORIO DE DATOS (EDA)

Análisis bivariado y multivariado





Solo existe un alto grado de correlación entre altura y peso. Además no vemos diferencias entre medallistas y no medallistas. Otro insight observable es como disminuye el rango de peso y altura en deportistas de mayor edad.

DESARROLLO DE UN MODELO DE ML

Modelos de clasificación

Se entrenó una gran cantidad de modelos utilizando:

- Segmentación por disciplinas
- Targets variados: 'Medal', 'Posición', 'Medal_Bool'.
- Modelos de regresión y clasificación.

Finalmente se utilizaron los siguientes modelos de clasificación utilizando 'Medal_Bool' como la variable target:

- Decision Tree Classifier
- Random Forest Classifier
- Gradient Boosting Classifier y sus variantes Light GBM y XGBoost
- Logistic Regression
- Naive Bayes

02 Proceso

ENTRENAMIENTO Y EVALUACIONES

RESULTADOS

03

MODELO FINAL

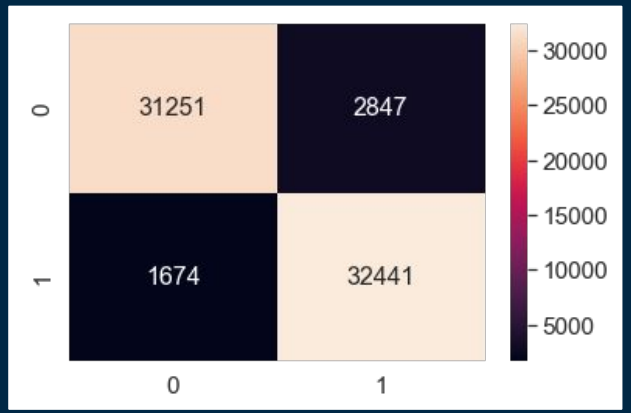
XGBoost

El modelo que mejores resultados obtuvo fue el XGB, habiendo obtenido los mejores parámetros con Hyperopt y resuelto el dataset desbalanceado con SMOTE.

Classification report

	Precision	Recall	F1-score	Support
No medal	0.95	0.92	0.93	34098
Medal	0.92	0.95	0.93	34115
Accuracy			0.93	68213

Matriz de confusión



CONCLUSIONES

- Los **principales determinantes** de la obtención de una medalla no son la altura, el peso o la edad. Sino el **país** que representan, si es un deporte de **equipo** o no y la cantidad de participantes.
- No es posible predecir el resultado solo en base a parámetros antropométricos ya que a este nivel, **los atletas de una misma disciplina tienen morfologías similares**.
- El **esfuerzo**, la **dedicación** y el **talento** de cada atleta quedan por fuera del alcance de este análisis

Pablo Tomás Fernández
tomasferc33@gmail.com

Pedro del Campo
pedrodelcampo123@gmail.com

GRACIAS