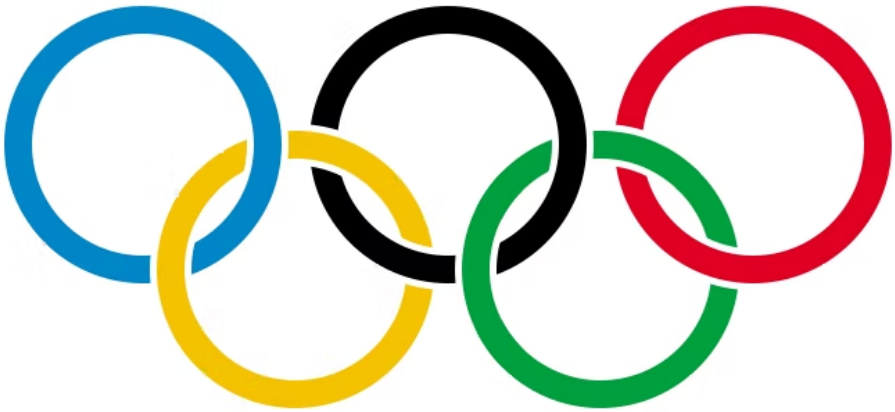
CODER HOUSE
PABLO TOMÁS FERNÁNDEZ - PEDRO del CAMPO

PROYECTO DATA SCIENCE

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



Índice

1. Abstract

Los Juegos Olímpicos modernos de verano se realizan cada cuatro años desde 1896. En ellos participan los mejores atletas en las principales disciplinas deportivas. El dataset que se analizó, contiene los datos de cada deportista individual incluyendo: sexo, edad, peso, altura, país, disciplina y edición en la que compitió y qué medalla ganó. A partir del mismo, es el objetivo de este proyecto analizar la relación entre estas distintas variables y cómo se modifican con el paso del tiempo. Es también el objetivo, desarrollar herramientas de Machine Learning para intentar determinar qué parámetros son determinantes para lograr el éxito olímpico.

Creemos que esta información puede ser beneficiosa para atletas, entrenadores, nutricionistas, directores deportivos y todo aquel que esté interesado en ganar una competencia deportiva. A partir de este análisis, se podrían establecer parámetros ideales a los cuales debería llegar cada atleta para tener más posibilidad de ganar la competencia. También es útil para reclutadores, para que tengan bien parametrizado qué tan probable es que un joven atleta sea capaz de ganar en un futuro.

2. Definición del objetivo

- Detectar cuáles son los pesos y altura óptimas o promedio de los participantes que han obtenido medallas, de igual manera la edad de los mismos.
- Segmentar el análisis en algunos deportes individuales y por el sexo de los deportistas.
- Detectar otros factores que afecten la obtención de medallas.

3. Contexto Comercial

Es común observar cómo la medición de la masa corporal (peso), forma parte de la rutina diaria del control biomédico del entrenamiento, debido a que es la variable morfológica más sensible a las cargas de trabajo, a los problemas de salud y del estado nutricional. Además, en algunos deportes es el criterio de selección de la modalidad o categoría deportiva en la cual debe participar el deportista.

La composición corporal juega un papel fundamental en el rendimiento del atleta. Además de permitir mejorar aspectos como la velocidad y la agilidad, reduce notablemente el riesgo de lesión, sobre todo a nivel articular.

Existen varios métodos de clasificación que dependen del deporte y de la Federaciones Internacionales (FI) de cada deporte, como podría ser la edad de los mismos.

4. Problema Comercial

El problema comercial sería poder identificar aquellos deportistas que cubren los valores óptimos de la disciplina, para que puedan participar en los JJOO

Ante ello, el objetivo de este trabajo es poder construir un algoritmo de clasificación binaria que, ante distintas variables, este pueda predecir, en base a sus características, qué tan probable es que se obtenga una medalla.

5. Data acquisition

Se obtuvo un dataset de Kaggle con los datos que contiene los datos sobre todos los atletas que compitieron en los JJOO desde 1896 hasta 2016. El mismo se puede obtener en el siguiente link: https://www.kaggle.com/datasets/josephcheng123456/olympic-historical-dataset-from-olympediaorg?datasetId=2379197&select=Olympic_Games_Medal_Tally.csv . Asimismo esta se obtuvo mediante el webscraping de la web www.olympedia.org.

A partir de 5 tablas incluidas en el dataset se confeccionó una tabla final. Las columnas de la esta tabla (posterior al proceso de feature engineering) fueron las siguientes:

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_Norm	float	normaliza la posición utilizando la cantidad de participantes. Obteniendo un número entre 0 y 1. Campo obtenido por feature engineering.

6. Preguntas e hipótesis

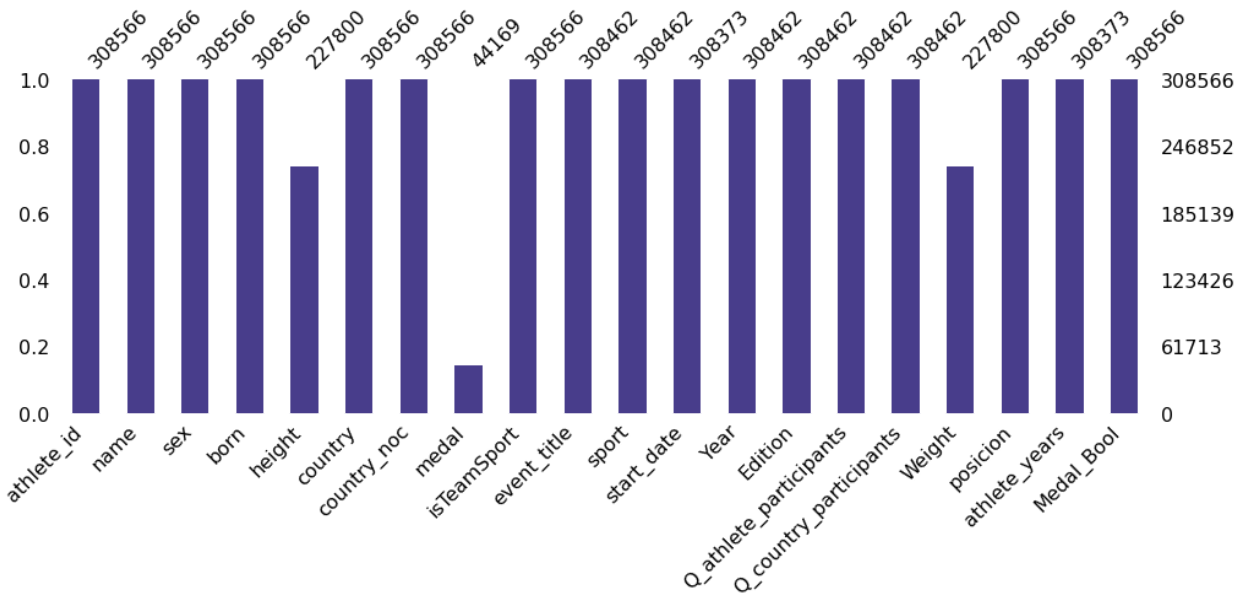
1. ¿Cómo se ha modificado el físico de los deportistas a lo largo del tiempo?
2. ¿Ha habido una mayor proporción de mujeres en los JJOO?
3. ¿En qué deportes es determinante la contextura física para obtener medallas?
4. ¿Cómo son las edades de los participantes en los distintos deportes?
5. ¿Es la edad un factor influyente al momento de ganar medallas?

Hipótesis:

- Hay deportes donde el peso y la altura son determinantes para el rendimiento deportivo.
- Hay edades en la cual los atletas suelen tener mejores resultados.
- Al pasar de los años se igualará el nivel de mujeres y hombres que participan en los JJOO.

7. Exploratory Data Analysis

Datos nulos

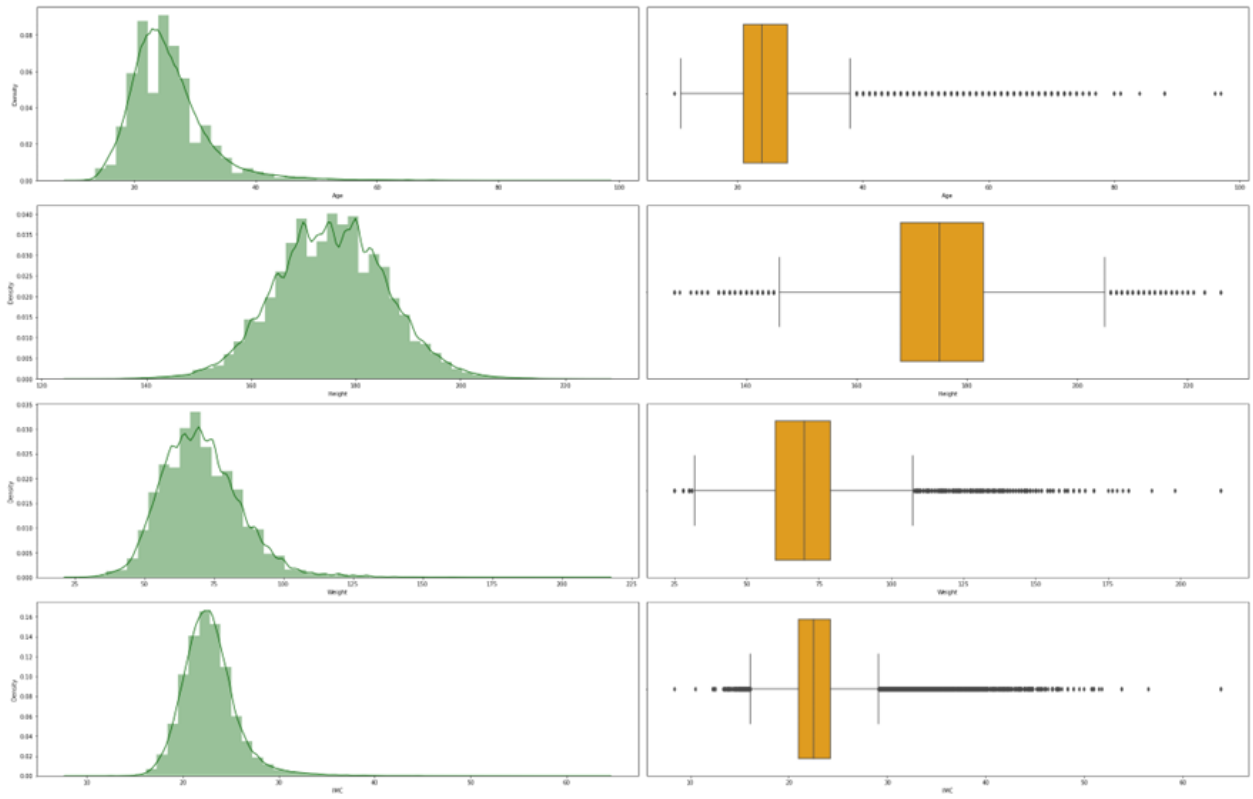


En el gráfico superior se visualizan las columnas y la cantidad de datos no nulos en cada una de ellas. Observamos que existen datos nulos en las columnas de peso y altura junto con medalla; lo cual es lógico ya que no todos los atletas obtienen medalla. Sin embargo, los datos faltantes en las variables antropométricas requerirán posterior manejo.

Análisis Univariado

En el siguiente apartado se explorará cada variable en particular, su distribución y los valores más prevalentes.

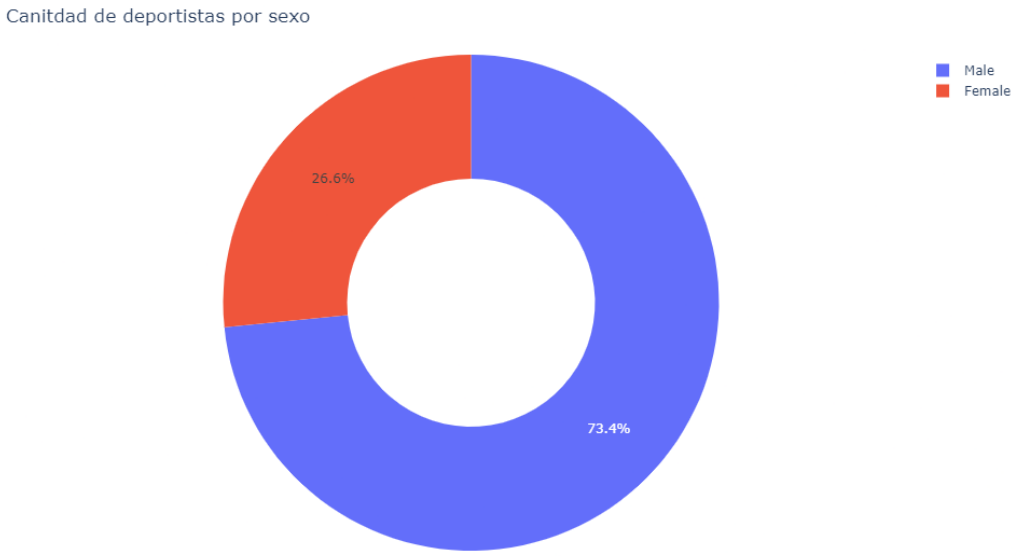
Variables continuas



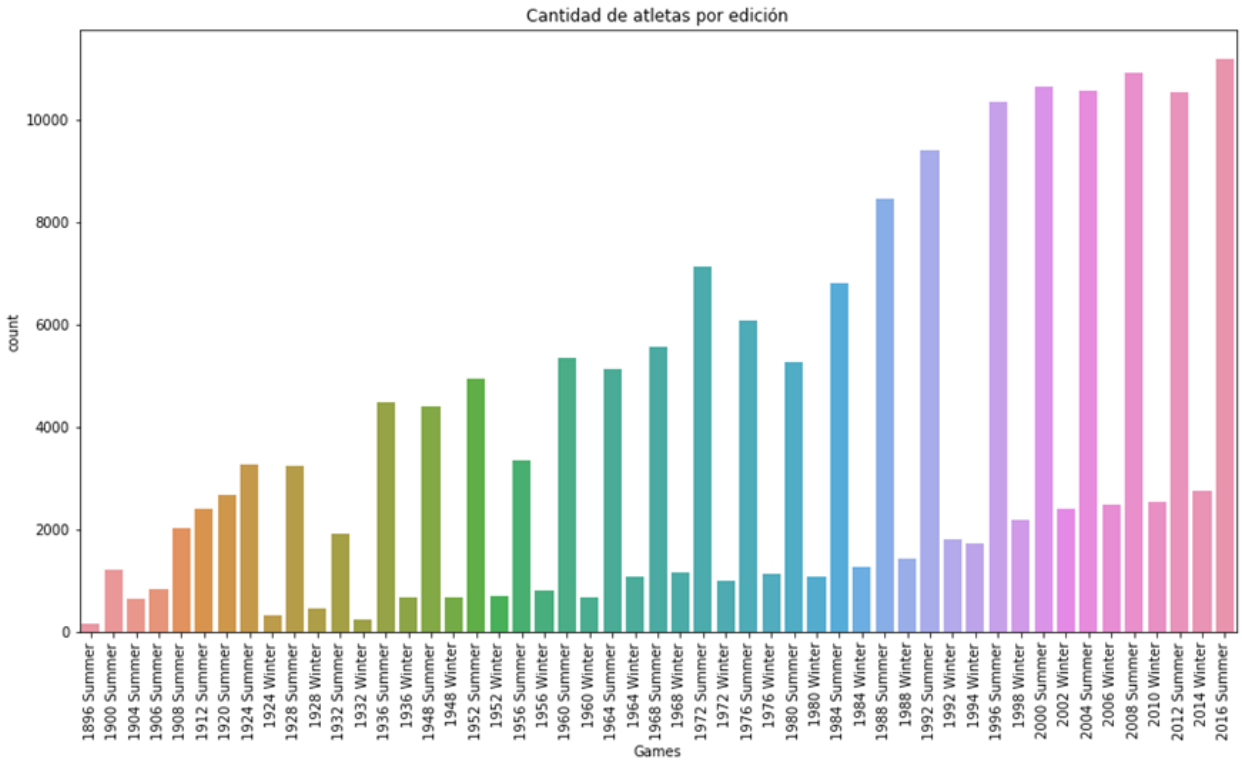
En los cuatro gráficos se puede observar la distribución de las variables continuas: edad, peso, altura e IMC. Las cuatro presentan una distribución gausiana en la cual la media, moda y mediana coinciden. En los cuatro gráficos son visibles numerosos outliers, sin embargo no se deben a errores de input por lo que decidimos mantenerlos.

En cuanto a la edad, vemos que la mediana está en 24 años, incluyendo atletas entre 10 y ¡97 años! El peso se sitúa entre un mínimo de 25 y un máximo de 214kg, con un promedio de 70kg. La altura va de 127 a 226cm con una media de 175cm. La gran variedad en estos valores se debe a la gran variedad de deportes que se incluyen en los JJOO, con sus diferentes morfologías corporales. Lo mismo sucede con la edad, teniendo una mayoría de atletas jóvenes pero incluyendo niños y adultos mayores.

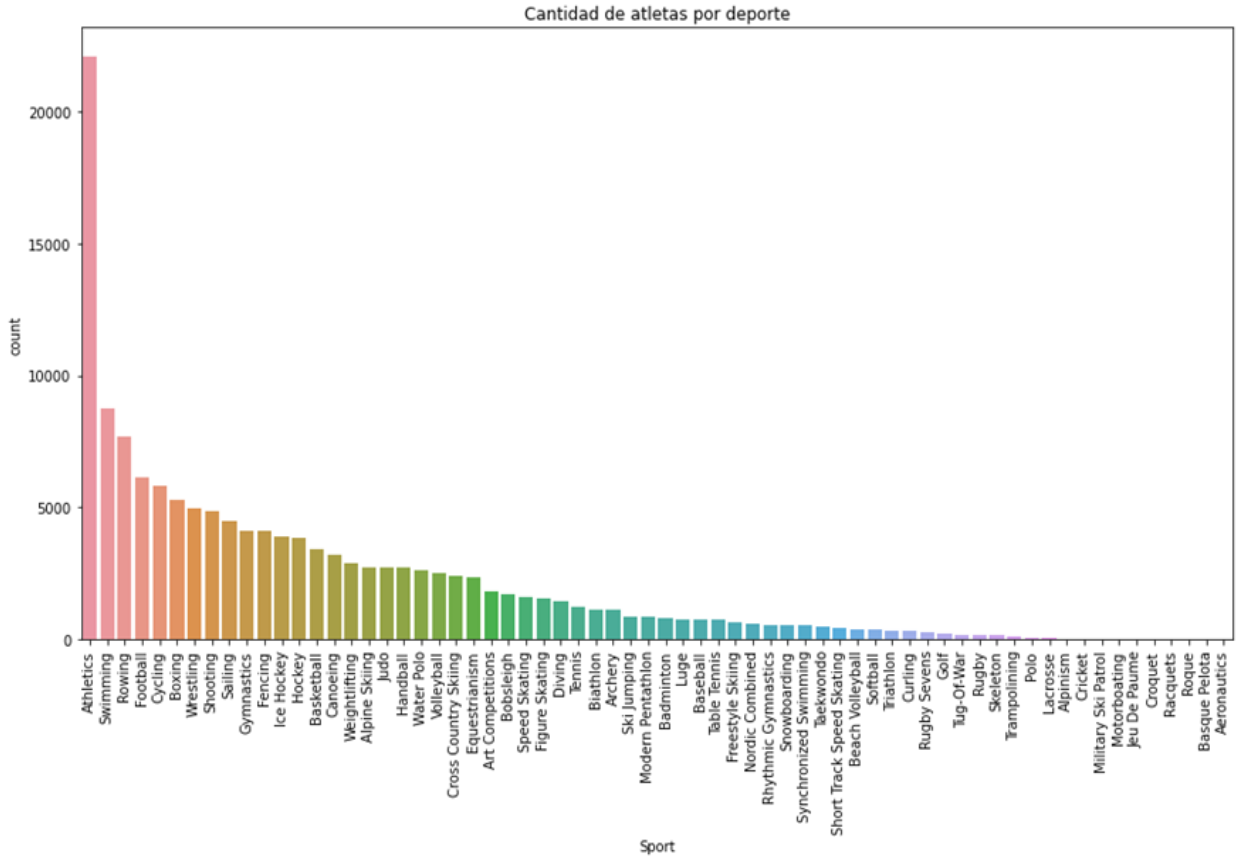
Variables categóricas



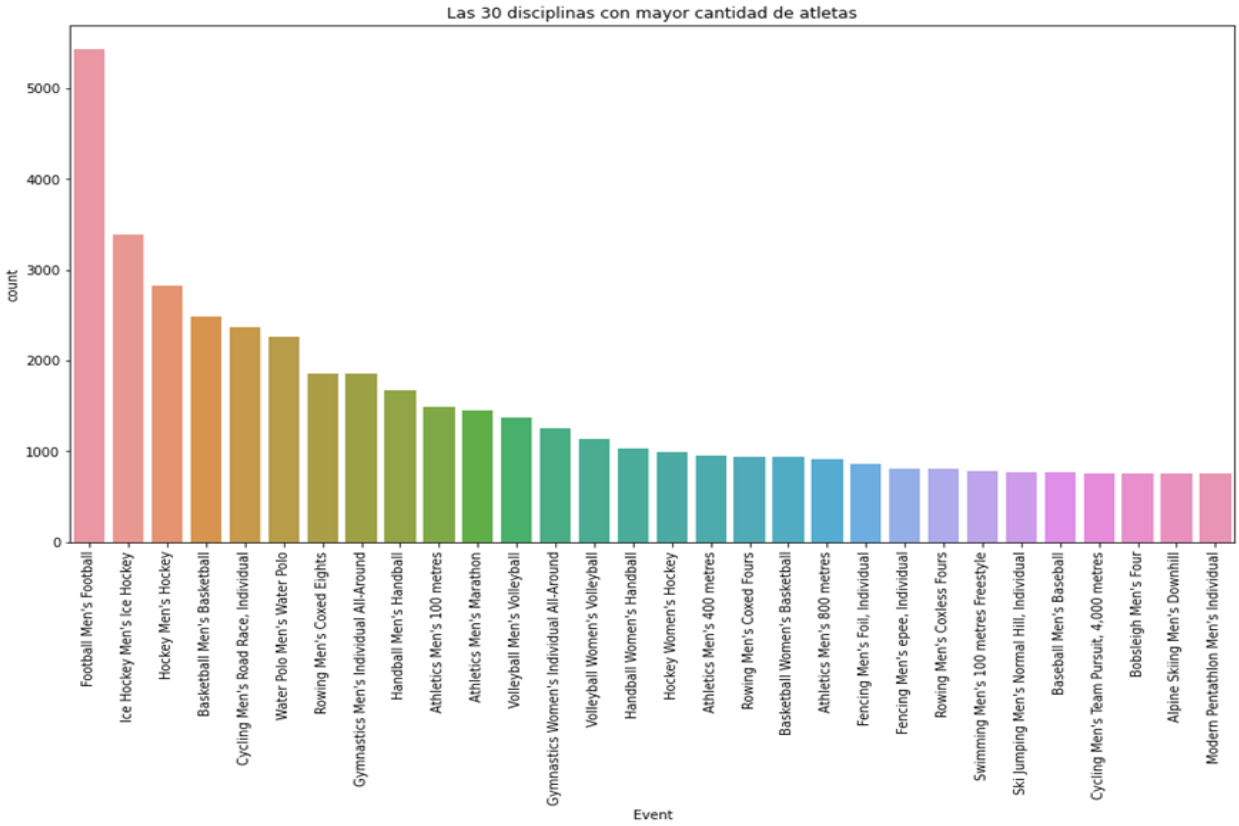
La cantidad de atletas masculinos supera tres veces la cantidad de atletas femeninas. Hay que tener en cuenta que esta proporción varía en las distintas disciplinas y con el tiempo la relación entre ambos sexos se ha vuelto más par.



La cantidad de atletas ha aumentado desde su inicio, tanto en invierno como en verano.

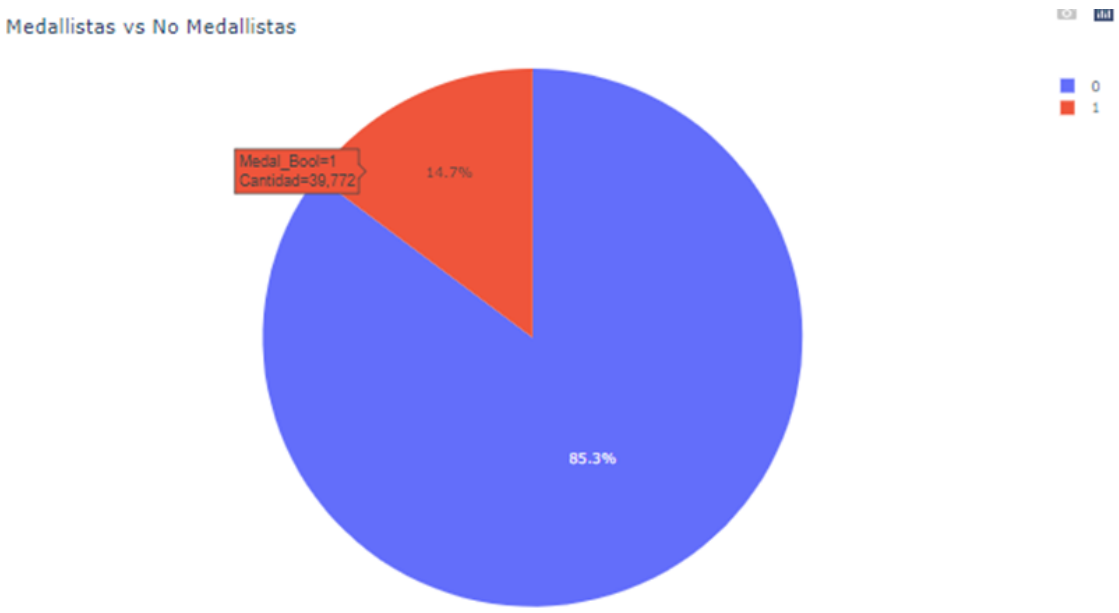


El deporte con más competidores es el Atletismo, seguido de Natación, Remo, Fútbol y Ciclismo



Las disciplinas con mayor cantidad de atletas son principalmente deportes de equipo: Hockey sobre hielo, Fútbol, Hockey sobre césped, Básquet, Waterpolo. Nótese como algunos deportes influyen muchas disciplinas y otros solo dos o tres.

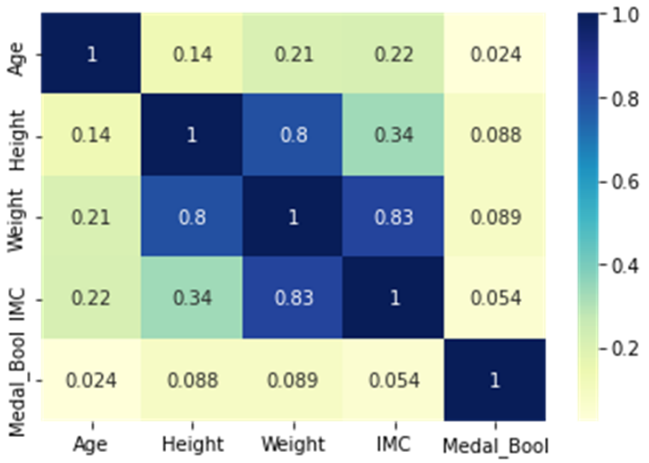
La gran mayoría de los atletas no obtuvieron una medalla y, ya que esta es nuestra variable target, debemos tener en cuenta que nuestro dataset está desbalanceado.

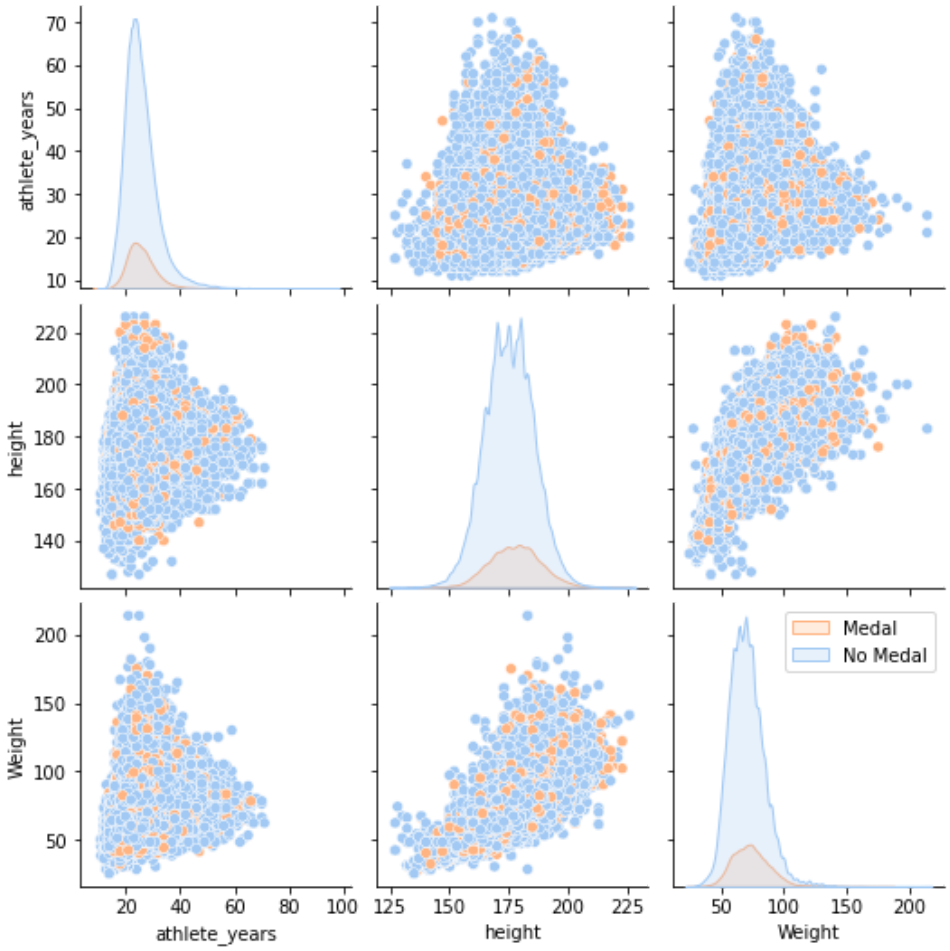


La cantidad de medallas de oro, plata y bronce son similares, lo cual tiene sentido ya que generalmente se entrega una de cada categoría y en algunos casos de empates se suelen entregar dos de una misma categoría.

Análisis bivariado

Variables continuas - Variables continuas



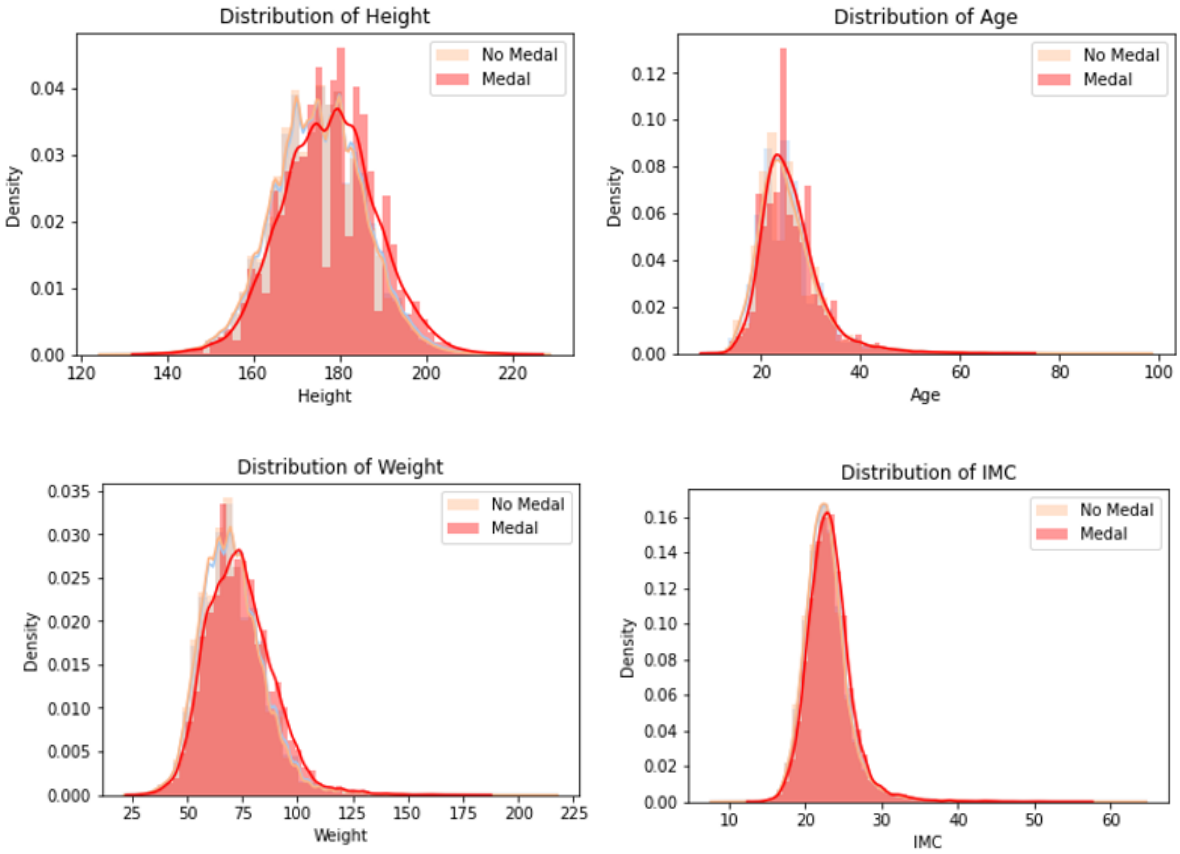


A partir del correlation map y el pairplot podemos ver como hay una alta correlación lineal entre peso y altura (83%). También entre el IMC y el peso, lo cual es lógico ya que el primero se calcula a partir del segundo.

Otro punto interesante es que los tamaños de los deportistas (peso, altura e IMC), suelen ser más variados en los atletas jóvenes que entre los adultos. Además podemos ver que no existen distribuciones diferentes para medallistas y no medallistas en el pairplot, por lo menos no a simple vista.

Variables continuas - Variables categóricas

Se evaluaron las correlaciones entre las diversas variables continuas y la principal variable categórica de importancia: Medal_Bool. Se graficaron histogramas y boxplots pero no se identificaron diferencias significativas en ambos grupos.



8. Data Wrangling

En esta etapa se realizaron distintas modificaciones a la tabla inicial.

Manejo de valores nulos

Con respecto al campo “medal”, se reemplazaron los nulos por el valor “N_Medal”. Los registros con campos nulos en otras columnas se eliminaron para facilitar el análisis.

One Hot Encoding y Label Encoding

Con el objetivo de optimizar el rendimiento de los modelos se realizaron procesos de One Hot Encoding sobre los campos con dos categorías: “isTeamSport”, “Edition” y “sex”; creando columnas de tipo booleano. Para las columnas con más categorías (“sport”, “country_noc” y “event_title”) se utilizó Label Encoding con el mismo objetivo.

Manejo de outliers

En esta sección se utilizó la técnica de Isolation Forest que consiste en seleccionar los outliers para luego eliminarlos y disminuir el ruido de estos valores.

Interval Cut and LabelEncoder

En esta sección se tomaron las variables continuas y se las separó en 10 grupos de la misma cantidad de atletas cada uno. Luego se aplicó Label Encoding sobre estos campos. Es decir, construimos campos de naturaleza categórica a partir de campos de naturaleza continua.

Normalización y estandarización

Se crearon columnas con datos estandarizados y normalizados a partir de todas las variables continuas. La variable “posicion” fue normalizada por evento y por edición. El resto fue hecho con todo el dataset

9. Desarrollo de modelos de ML

En esta etapa se intentará desarrollar un modelo de Machine Learning eficaz para predecir el rendimiento de los atletas en base a los datos proveídos. Las variables target variarán entre: “posicion_Norm”, “Medal_bool” y “Medal”. Y para entrenar el modelo se utilizarán todas las otras variables.

Modelos de regresión

En un primer momento, intentaremos desarrollar un modelo de regresión, utilizando como variable objetivo la posición obtenida (previamente normalizada con la cantidad de participantes). Para lo siguiente utilizaremos: regresión lineal, con sus variantes de Ridge y Lasso, árboles de decisión, random forests y Gradient Boosting.

En esta sección se evaluó el dataset entero y los resultados fueron muy malos. Las métricas utilizadas -R cuadrado y RMSE- eran muy diferentes a las esperadas.

```
Random Forest Test RMSE :  0.2581004014601578
Random Forest Test R-squared :  -0.21400710153290037
Decision Tree Test RMSE :  0.3529760977337828
Decision Tree Test R-squared :  -1.2705677348015905
Linear Regression Test RMSE :  0.23078798081544233
Linear Regression Test R-squared :  0.02933305136145048
Lasso Test RMSE :  0.23434575569758143
Lasso Test R-squared :  -0.00082478442882760 74
Ridge Test RMSE :  0.23078806842213484
Ridge Test R-squared :  0.02933231443453821
Gradient Boosting Test RMSE :  0.23154684533999656
Gradient Boosting Test R-squared :  0.022939167748364575
```

Luego se entrenaron modelos para cada disciplina. Para la gran mayoría de estos modelos las métricas resultaron muy desalentadoras. En aquellas disciplinas en las cuales las métricas eran aceptables pudimos detectar overfitting por la poca cantidad de atletas. Por esto, debimos intentar cambiar la forma de encarar el problema.

	event	model	R-squared	RMSE	Q_athletes
3043	Coxless Fours, Men1	Decision Tree	1.000000	0.000000e+00	9.0
3047	Coxless Fours, Men1	Gradient Boosting	1.000000	4.249697e-07	9.0
3190	Middleweight (≤73 kilograms), Men	Ridge	0.986764	1.232633e-02	10.0

Modelos de clasificación

Para esta etapa se entrenaron diversos modelos de clasificación: regresión logística, árboles de decisión, Random Forests; y modelos de ensamble: Gradient Boosting, con sus variantes XGB y LGB.

Modelos de clasificación por disciplina

En esta sección se obtuvieron resultados mínimamente mejores. Pero aun así con mucho overfitting dado la pequeña muestra que determina cada categoría “Event”. Por lo tanto, se modificó a un modelo entrenándolos como un solo conjunto de datos

Modelos de clasificación general - Primera evaluación

A partir de esta instancia, se utilizará el dataset completo para entrenar los modelos y continuará de esta manera hasta el final del proyecto. Para iniciar intentamos observar qué variante de las variables era la mejor. Es decir, si convenía utilizar los datos estandarizados, normalizados, separados por rangos o puros.

Evaluación según el formato de los datos

Se entrenaron todos los modelos con las 4 variantes de datos y se recopilaron las métricas de evaluación en una tabla para facilitar la comparación. En esta tabla figuran el modelo utilizado, la modalidad de los datos, la precisión, F1-score, recall y accuracy. Las mejores métricas fueron de los datos separados en rangos. A partir de esta etapa, utilizaremos este formato de datos para entrenar a los modelos.

model	Accuracy	Precision	Recall	F1-Score	Tipo
XGBClassifier	0.880601	0.802365	0.303757	0.440682	Datos_Rangos
XGBClassifier	0.879289	0.797840	0.295284	0.431039	Datos_Puros
XGBClassifier	0.879289	0.797840	0.295284	0.431039	Datos_Norm
XGBClassifier	0.879289	0.797840	0.295284	0.431039	Datos_Std

Validación cruzada

En esta etapa, se utilizó el método de validación cruzada para obtener métricas más sólidas. Esto se realizó mediante el algoritmo de StratifiedKFolds. Alternar los sets de entrenamiento y evaluación permite homogeneizar las métricas y sin obtener resultados aislados.

Los mejores promedios del F1-score fueron:

- DecisionTree de 0.55,
- RandomForest de 0.52
- XGBClassifier con 0.43

El accuracy es bastante alto en la mayoría de los modelos. Esto se debe a que el modelo está desbalanceado (la mayoría de los deportistas no ganaron medalla). Los modelos tienden a predecir resultados negativos. Por lo tanto, la mayor parte del trabajo consistirá en aumentar las otras métricas sin disminuir considerablemente la accuracy.

Segunda evaluación de modelos

Optimización de hiperparámetros

En esta etapa se utilizará la librería Hyperopt con el objetivo de buscar los mejores hiperparámetros para los modelos que tuvieron una mejor performance en la evaluación anterior: Decision Trees, Random Forests y XGB. Como mencionamos antes, seguiremos utilizando los datos en formato rangos.

Aquí un ejemplo de los parámetros obtenidos para el modelo de árbol de decisión:

```
Eval 3 - Score: 0.4227873601059181
{'class_weight': 0, 'criterion': 0, 'max_depth': 25, 'max_features': 2, 'max_leaf_nodes': 85,
'min_impurity_decrease': 1.4711850505052699e-05, 'min_samples_leaf': 8,
'min_samples_split': 15}
```

Evaluación

Para evaluar estos modelos utilizaremos:

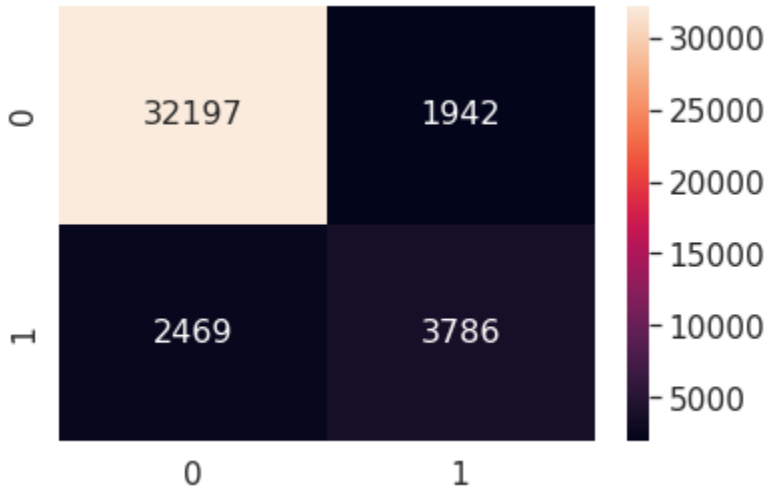
- Classification Report
- Matriz de confusión
- ROC-AUC Score
- Detección de overfitting

Observemos el ejemplo del RandomForestClassifier. En primer lugar aquí tenemos el reporte de clasificación. Observamos como la accuracy es alta pero sigue teniendo recall y precisión bajas para los que obtuvieron medalla. Aún así es una gran mejora con respecto a los modelos de la etapa anterior.

	precision	recall	f1-score	support
0	0.93	0.94	0.94	34139
1	0.66	0.61	0.63	6255
accuracy			0.89	40394
macro avg	0.79	0.77	0.78	40394
weighted avg	0.89	0.89	0.89	40394

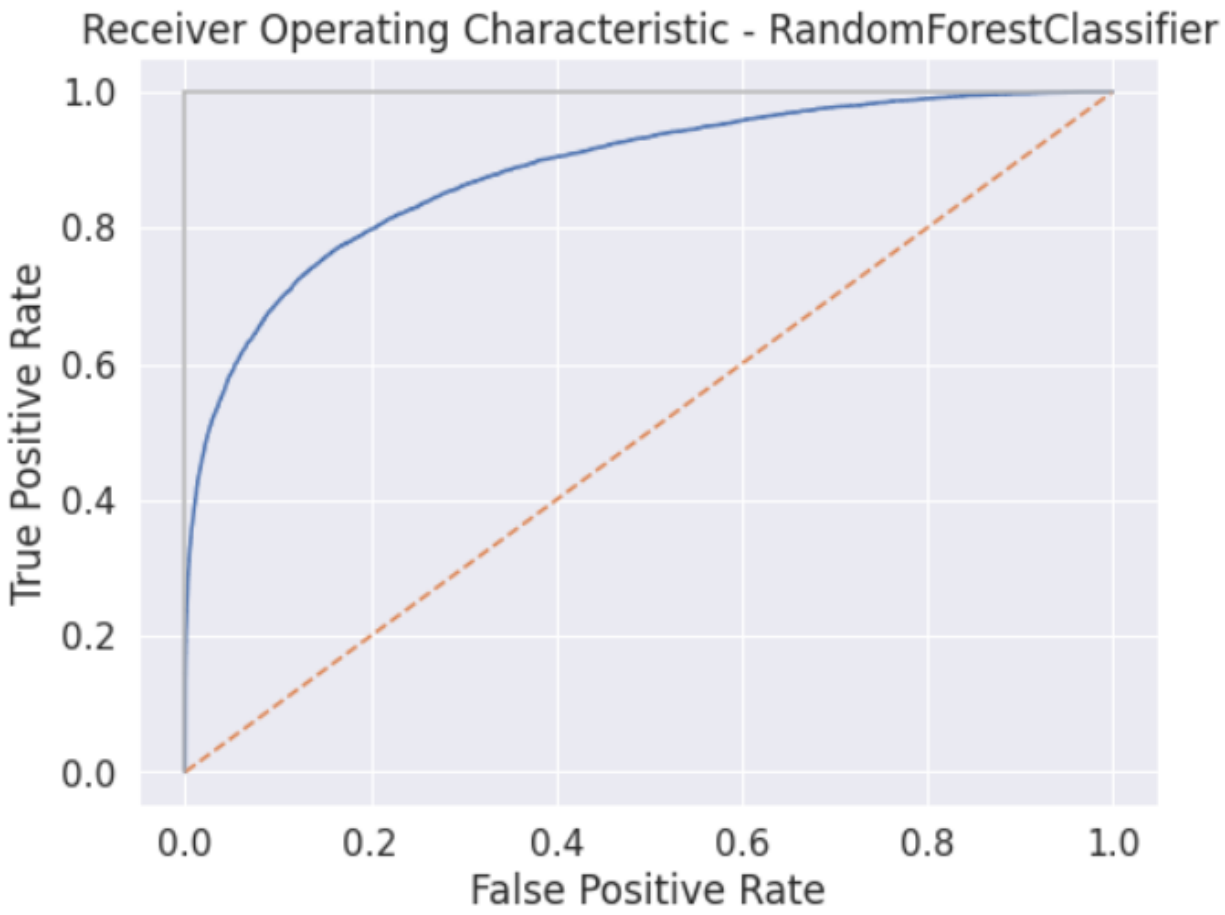
Accuracy: 0.8908006139525673

Luego aparece la matriz de confusión donde se visualiza el desbalance del dataset y se entiende las métricas del cuadro anterior.

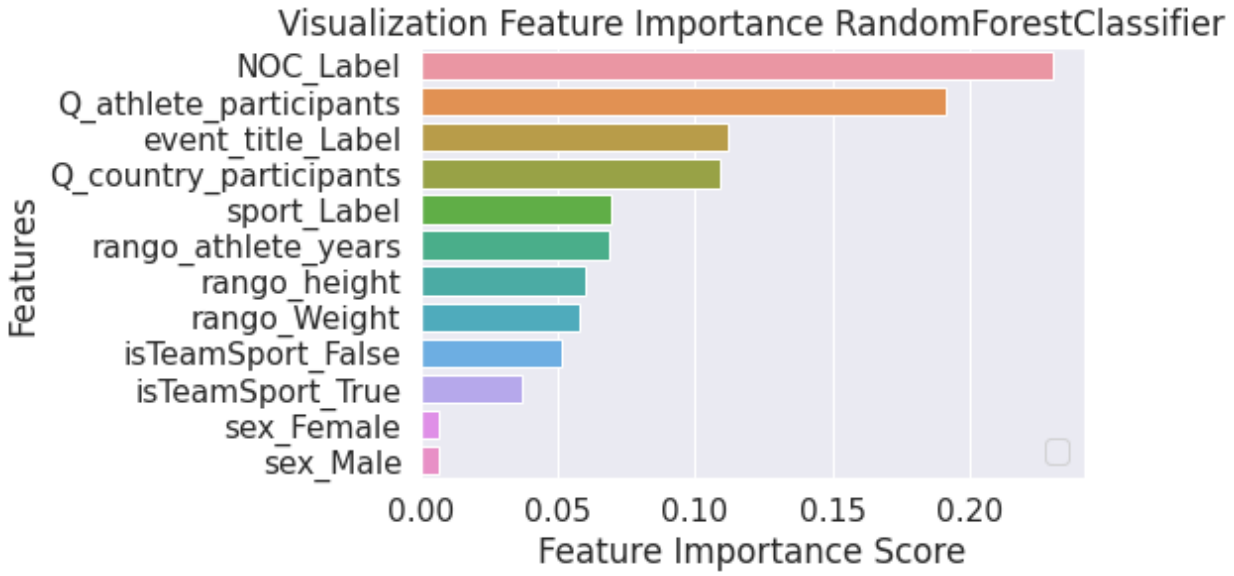


A continuación aparecen el ROC-AUC score y el gráfico correspondiente.

```
ROC AUC SCORE para la RandomForestClassifier:  0.8834328032462574
```



Luego se visualiza el Feature Importance del modelo mostrando la relevancia de cada columna en el funcionamiento del mismo. Es interesante observar como las variantes más importantes no son la edad o el peso o la altura sino el país (‘NOC_Label”), la cantidad de participantes (“Q_athlete_participants”) en la disciplina y qué disciplina es (“event_title_Label”).



Análisis de componentes principales (PCA)

En esta sección se realizó un PCA. Aquí se observan los resultados de la varianza acumulada por cada uno de los 6 componentes:

```
Component 1: 0.39
Component 2: 0.57
Component 3: 0.74
Component 4: 0.88
Component 5: 0.97
Component 6: 1.00
```

Se observa como el último componente es el único que no tiene mucha relevancia. Se volvieron a entrenar los modelos eliminando este último componente y se volvieron a evaluar. Los resultados fueron peores que con todos los componentes, por lo tanto se descartó utilizar esta selección de datos.

Tercera evaluación

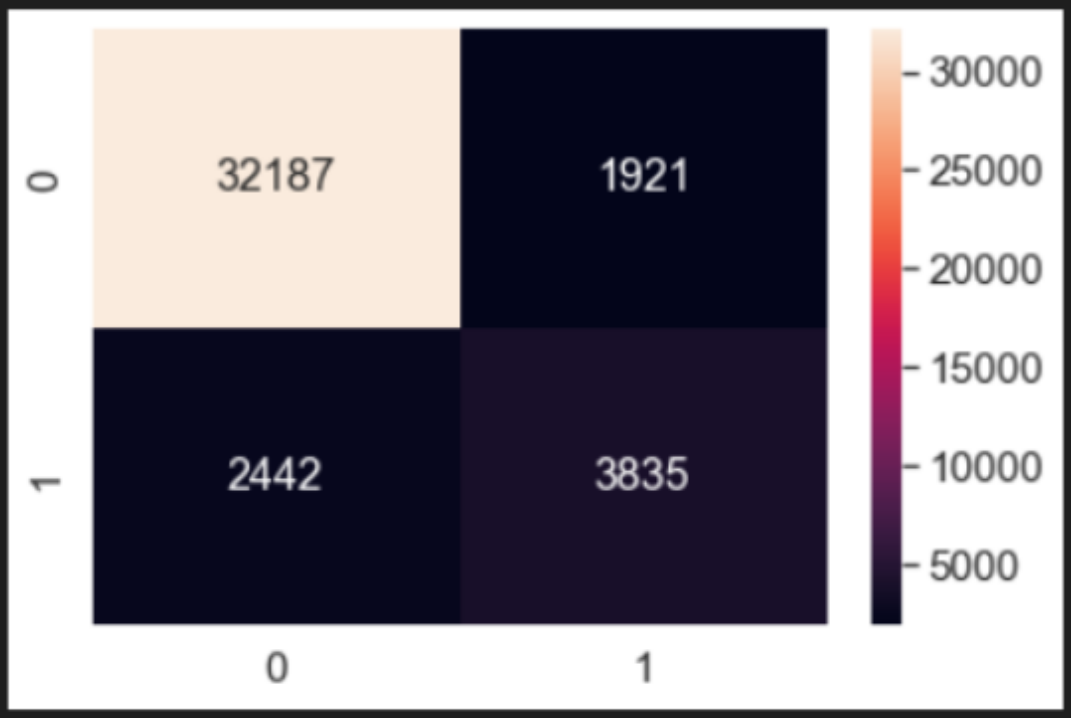
En esta última ronda de evaluación utilizaremos el modelo XGBoost que fue el que mejor performance tuvo hasta el momento. Aquí se volverá a evaluar utilizando los mejores hiperparámetros encontrados utilizando la técnica SMOTE y comparándola con el mismo

dataset sin modificar. La técnica SMOTE se utiliza para solucionar el problema que conlleva utilizar un dataset desbalanceado.

Dataset normal

Estos son los resultados para el dataset sin utilizar SMOTE:

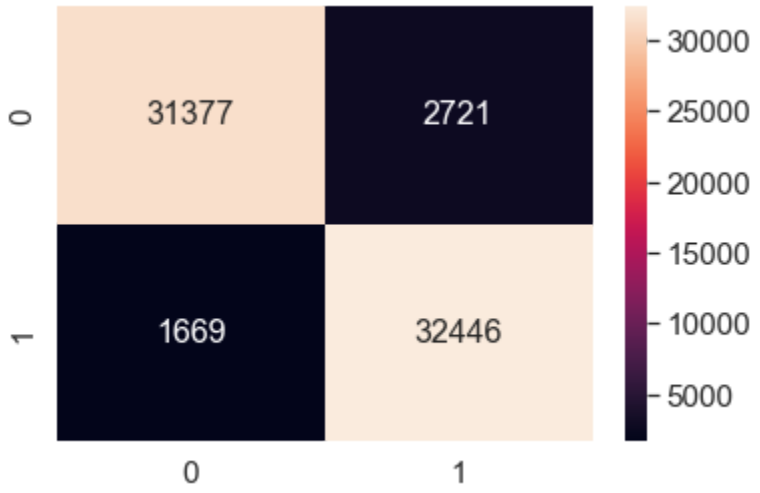
	precision	recall	f1-score	support
0	0.93	0.94	0.94	34108
1	0.67	0.61	0.64	6277
accuracy			0.89	40385
macro avg	0.80	0.78	0.79	40385
weighted avg	0.89	0.89	0.89	40385



Dataset utilizando SMOTE

"Oversampling" es una técnica que genera registros "artificiales" a partir de los existentes en la clase minoritaria para equilibrar las dos clases. Lo opuesto sucede con el "undersampling" que elimina registros de la clase mayoritaria con el mismo fin. Y la técnica SMOTE utiliza ambos para obtener un dataset balanceado. Observemos los resultados con está técnica.

	precision	recall	f1-score	support
0	0.95	0.92	0.93	34098
1	0.92	0.95	0.94	34115
accuracy			0.94	68213
macro avg	0.94	0.94	0.94	68213
weighted avg	0.94	0.94	0.94	68213



10. Conclusiones

Como conclusión principal podemos decir que pudimos entrenar satisfactoriamente un modelo de Extreme Gradient Boosting, optimizando los hiperparámetros y balanceando el dataset utilizando SMOTE.

- Los factores que más contribuyeron a la predicción no fueron los esperados.
- A pesar de contar con un dataset muy extenso, es muy heterogéneo, con variedad de deportes, contexturas físicas que además fueron variando con los años.
- No parece haber relaciones estrechas entre el peso, la altura, la edad y la obtención de medallas.

Entre los factores que más contribuyen a las predicciones se encuentran: el país que representan, si es un deporte de equipo o no y la cantidad de atletas que participan.

¿Cómo se explica esto?

Tomemos el ejemplo del básquetbol. Este es un deporte dominado por Estados Unidos. Solo 3 veces desde 1936 no ganaron la medalla de oro, y en esos casos obtuvieron medallas de plata y bronce. Por lo tanto, si un atleta, es basquetbolista (deporte de equipo) y es estadounidense, es altamente probable que gane la medalla dorada.

Esta situación se repite en varios deportes: Jamaica en velocidad, Cuba en boxeo, Fiji en rugby, y muchos más. Esta es la razón por la cual el país es más importante que el peso, la altura y la edad.

Otro campos que no determina nada es el sexo. La gran mayoría de las disciplinas están separadas por sexo lo que anula las diferencias. Y en aquellas pocas en las que es un equipo mixto donde no parece tener inferencia.

Aun habiendo obtenido un modelo que predijo adecuadamente los datos, podemos obtener un aprendizaje importante a partir de esto: los atletas ganadores son más que solo su cuerpo. No es la edad, ni la altura, ni el peso por si solos lo que determina el éxito. Deberíamos obtener datos sobre las horas de entrenamiento, las horas de análisis de video, las horas de kinesiología y viajes que cada atleta realiza para tener un panorama más completo sobre la formación de los campeones.

11. Futuras líneas de exploración

Este proyecto abre la puerta a muchas investigaciones futuras. Sería interesante relevar datos sobre el entrenamiento de los atletas, el tiempo que le dedican a cada componente del mismo (musculación, técnica, estrategia); sobre su salud, sus lesiones, intervenciones quirúrgicas y los profesionales que los atienden; el equipo que tienen detrás, la cantidad de los profesionales, kinesiólogos, entrenadores, preparadores físicos; los recursos económicos que intervienen, apoyo estatal, sponsors, premios. Estos son solo algunos de los factores que se podrían evaluar y que son modificables (a diferencia del país que representan).

12. Autores

Pablo Tomás Fernández
tomasferc33@gmail.com

Pedro del Campo
pedrodelcampo123@gmail.com

Tutor: Alfredo Parente