

Predicción de Medallas Olímpicas

Análisis y Modelado



Medallas



Análisis



Datos



Modelado

Hecho por Daniel Gatón, Maria Velic, Eric García

Introducción y Objetivos

El objetivo principal de este proyecto es predecir la probabilidad de que un atleta obtenga una medalla olímpica, basándose en sus características demográficas y físicas, así como en el país de origen.

Objetivos

- Análisis exploratorio de datos para comprender la estructura y patrones
- Preprocesamiento de datos para manejar valores nulos y desequilibrio de clases
- Desarrollo de un modelo predictivo utilizando Random Forest
- Validación del modelo y análisis de resultados

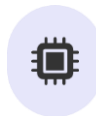
Relevancia

- Detección de talento y optimización de entrenamientos deportivos
- Desarrollo de sistemas de recomendación y entrenadores virtuales
- Análisis de factores que influyen en las oportunidades de éxito olímpico
- Investigación en inteligencia artificial aplicada al deporte



Deportivo

Identificación de talento y estrategia de entrenamiento



Tecnológico

Sistemas de recomendación y entrenadores virtuales inteligentes



Investigación

Análisis de género y nacionalidad en oportunidades de éxito

Dataset y Definición del Problema



Dataset

- Nombre:** "120 years of Olympic history: athletes and results"
- Estructura:** 15 columnas y 271.116 filas
- Representación:** Participaciones individuales en los Juegos Olímpicos
- Contenido:** Datos demográficos, físicos y resultados de atletas

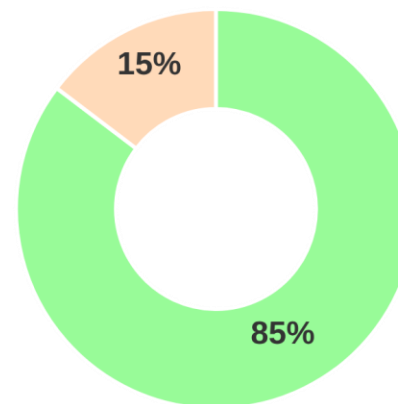


Definición del Problema

- Tipo:** Clasificación binaria
- Objetivo:** Predecir si un atleta ganará una medalla olímpica
- Variable Target:** "Medaled" (1: sí, 0: no)
- Desafío:** Desequilibrio de clases (85,3% no medallistas vs. 14,7% medallistas)



Distribución de la Variable Target



Medallistas
39.783 (14,7%)



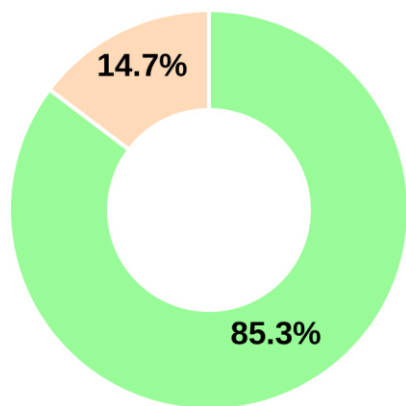
No Medallistas
231.333 (85,3%)

Análisis Exploratorio: Desequilibrio y Valores Nulos

El análisis exploratorio reveló desafíos significativos en el conjunto de datos que afectarán el desarrollo del modelo predictivo.



Desequilibrio de Clases

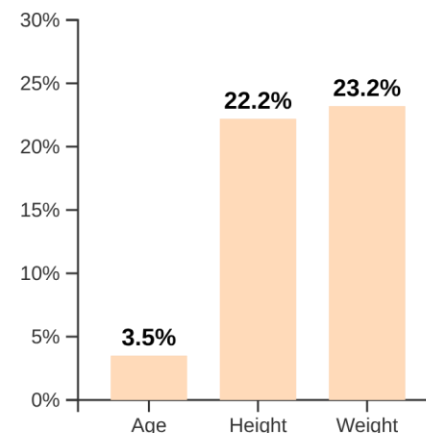


La variable objetivo "**Medaled**" presenta un marcado desequilibrio:

■ No medalla: **85,3%** (231.333) ■ Medalla: **14,7%** (39.783)



Valores Nulos



Se identificaron valores nulos en varias columnas clave:

3,5%

Age

22,2%

Height

23,2%

Weight






Implicación: El desequilibrio de clases y la alta proporción de valores nulos requerirán estrategias específicas de manejo para el entrenamiento del modelo.

Análisis de Variables Numéricas



Estadísticas Descriptivas

Variable	Media (aprox.)	Mediana (aprox.)	Desviación Estándar (aprox.)
 Age	25,6 años	25 años	6,4 años
 Height	175,3 cm	175,0 cm	10,5 cm
 Weight	70,7 kg	70,0 kg	14,3 kg



Edad

La mediana de edad de los ganadores de medalla es ligeramente superior a la de los no ganadores, sugiriendo que la experiencia puede ser un factor predictivo.



Altura

No existe una altura "óptima" universal, ya que la relevancia varía significativamente según el deporte y la posición.



Peso

Similar a la altura, no hay un peso "óptimo" universal. La relevancia de este factor varía considerablemente entre diferentes disciplinas deportivas.



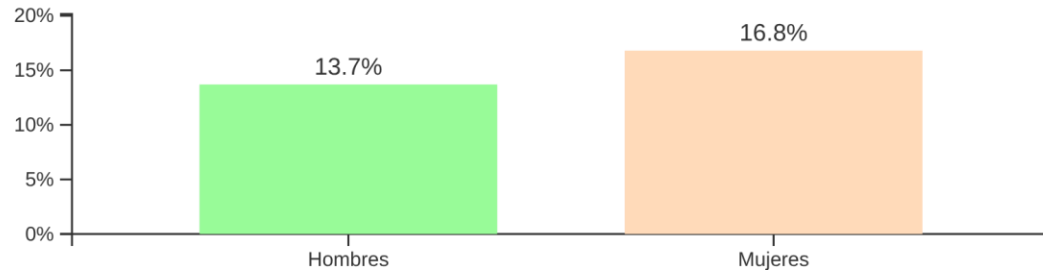
Conclusión: Experiencia (edad) y ciertas características físicas pueden ser factores predictivos relevantes, aunque su relevancia varía significativamente según el deporte.

Análisis de Variables Categóricas

El análisis de las variables categóricas 'Sexo' y 'País de Origen' revela su fuerte impacto predictivo en la obtención de medallas olímpicas.



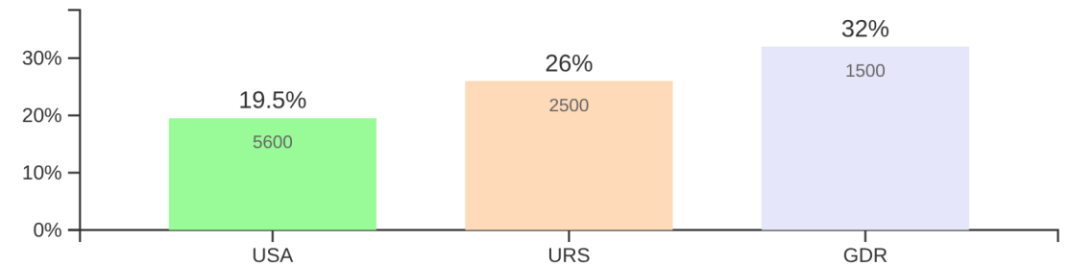
Análisis por Sexo



- Tasa de éxito de hombres: **13,7%**
- Tasa de éxito de mujeres: **16,8%**
- Las mujeres presentan una tasa de éxito ligeramente superior a la de los hombres.
- Puede atribuirse a la menor variedad de eventos de participación femenina en los primeros años del historial olímpico.



Análisis por País de Origen



- El país de origen es un predictor extremadamente fuerte de éxito.
- Países con políticas deportivas intensivas muestran tasas de éxito más altas.
- Ejemplo: URS (Unión Soviética) con 26,0% de tasa de éxito.
- Estados Unidos (USA) tiene más participaciones totales pero una tasa de éxito menor (19,5%).



Conclusión: Ambas variables categóricas demuestran tener un fuerte poder predictivo para la obtención de medallas olímpicas, con el país de origen siendo particularmente influyente.

Preprocesamiento de Datos

El preprocesamiento de los datos fue fundamental para preparar el conjunto de datos para el entrenamiento del modelo, abordando principalmente la gestión de valores nulos y la creación de la variable objetivo.



Creación de 'Medaled'

- Transformación de la columna original 'Medal'
- Reemplazo de valores NaN con 'NoMedal'
- Creación de la columna binaria 'Medaled'
- Asignación: 1 para medallistas, 0 para no medallistas



Imputación de 'Age'

- Se detectaron 9.484 valores faltantes (3,5%)
- Se aplicó imputación global
- Se rellenaron con la mediana de la columna 'Age' completa



Imputación de 'Height' y 'Weight'

- 'Height': 60.171 valores faltantes (22,2%)
- 'Weight': 62.875 valores faltantes (23,2%)
- Se optó por una estrategia de imputación segmentada
- Se rellenaron con la mediana calculada por deporte



Resultados del Preprocesamiento

- Conjunto de datos completo y limpio
- Variables numéricas imputadas correctamente
- Variable objetivo binaria correctamente definida
- Preparación exitosa para el entrenamiento del modelo

Modelado y Evaluación



Modelo Random Forest

- Se empleó un modelo Random Forest con 200 árboles
- División del conjunto de datos: 70% entrenamiento y 30% prueba
- Se mantuvo la proporción de clases (14,7% medallas vs 85,3% no medallas) en ambos conjuntos
- Semilla aleatoria (RSEED = 42) para reproducibilidad
- Pipeline de scikit-learn para automatizar transformaciones



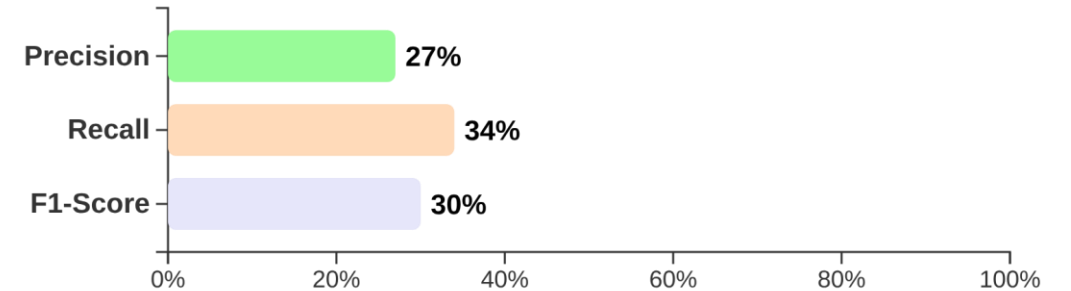
Limitaciones del Modelo

El modelo actúa como un "clasificador perezoso", inclinándose hacia la clase mayoritaria de no ganadores y fallando significativamente en la identificación correcta de los medallistas.



Evaluación del Rendimiento

Métricas obtenidas para la predicción de medallas olímpicas



Interpretación

- Bajas métricas indican un rendimiento insatisfactorio
- Precision (0,27): pocas medallistas correctamente identificadas
- Recall (0,34): capacidad limitada para encontrar a los medallistas
- F1-Score (0,30): puntuación F1 promedio

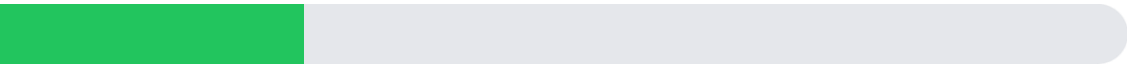
Conclusiones y Limitaciones



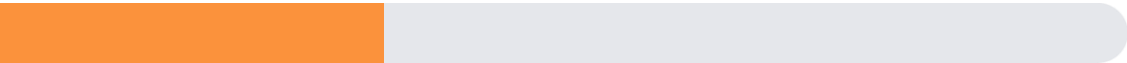
Resultados

El modelo Random Forest desarrollado no ha demostrado ser eficaz para predecir la obtención de medallas olímpicas.

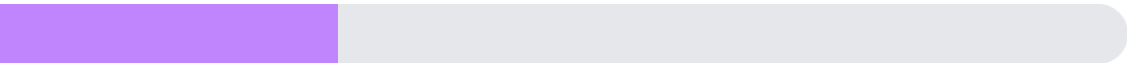
Precision: 0,27



Recall: 0,34



F1-Score: 0,30



Comportamiento del Modelo

- El modelo actúa como un "clasificador perezoso"
- Inclina significativamente hacia la clase mayoritaria de no ganadores
- Alto número de falsos positivos y falsos negativos
- Limitada utilidad práctica para la detección de talento o la planificación estratégica



Limitaciones

- Desequilibrio de clases: 85,3% de no medallistas vs. 14,7% de medallistas
- Poor performance en la identificación de la clase minoritaria (ganadores de medallas)
- Características demográficas y físicas limitan la capacidad predictiva
- Imputación de valores nulos en variables clave (Height, Weight)



Recomendaciones Futuras

- Probar algoritmos alternativos a Random Forest
- Recopilar más características relevantes para la predicción de medallas
- Mejorar la estrategia de manejo del desequilibrio de clases
- Reducir la cantidad de valores faltantes en variables clave
- Expandir el conjunto de datos con más recientes y relevantes features