Proyecto: Simulación de Eventos Discretos

María Fernanda Suárez González, Adrián Navarro Foya y Daniel Polanco Pérez

Grupo C311
https://github.com/MaferGlez03/Simulation-Swimming-Prediction

1. Introducción

La natación en los Juegos Olímpicos es uno de los eventos más destacados y populares, representando una mezcla perfecta de habilidad atlética, resistencia y estrategia. Se compone de 28 eventos individuales entre femeninos y masculinos, donde encontramos: los estilos libres (50 metros, 100 metros, 200 metros, 400 metros, 800metros y 1500 metros), mariposa (100 metros y 200 metros), espalda(100 metros y 200 metros), pecho(100 metros y 200 metros) y combinados (200 metros y 400 metros).

En víspera de los Juegos Olímpicos de París 2024 , el interés por conocer los resultados de esta disciplina aumenta significativamente. Es por ello que el objetivo principal de este proyecto es construir un modelo de predicción basado en el análisis de datos, capaz de predecir con buena precisión los tiempos de clasificación y las posibilidades de medalla para los atletas, así como los posibles podios para cada uno de los eventos. Para desarrollar esta predicción contamos con el nombre de los atletas, marcas de cada uno, fecha de cada marca, evento(sexo, distancia y estilo) en que compitió como variables de interés; por otro lado el nombre de cada competencia, país que representa cada atleta, y fecha de nacimiento como las variables que describen el problema.

2. Detalles de Implementación

Inicialmente, para obtener los datos necesarios, buscamos en la página oficial de la Federación Internacional de Deportes Acuáticos (World Aquatics) y descargamos por cada semestre desde enero de 2022 hasta la actualidad un csv por cada evento, que contienen el top 200 de las mejores marcas en dicho período.

Se importan las bibliotecas necesarias como numpy, pandas, sklearn para el KDE (Kernel Density Estimation), matplotlib, y seaborn para visualización. Una vez que se tienen los datos se hace uso de la función `parse_time_to_seconds` para convertir los tiempos de los atletas desde un formato de mm:ss a segundos totales, con la finalidad de facilitar así su manejo numérico en las simulaciones. Luego, cada una de estas marcas las multiplicamos por un valor que dependerá de un parámetro *alpha* y de la cantidad de marcas de ese atleta en ese evento. La idea de esta modificación a las marcas es 'premiar' de cierta manera a los atletas con más marcas y 'castigar' a los que tengan menos.

Utilizamos la Estimación de Densidad de Kernel (KDE), esto nos permite modelar la distribución de los tiempos de cada atleta en un evento de una manera "suave" y continua. Al estimar con KDE también le asignamos un peso a cada marca de un atleta el cuál estará dado por cuán reciente es esa marca, la asignación de un peso a cada marca se realiza a través de la función `date_to_value`. Posteriormente para cada atleta, se extrae una muestra de la distribución KDE generada, simulando así un tiempo de competición para ese atleta. Mediante la función `race_simulation`, se realizan simulaciones individuales de carreras donde se generan tiempos de competición para cada atleta a partir de sus distribuciones KDE.

Cada simulación de un evento consiste en 3 etapas: eliminatorias, semifinales y final (2 en el caso de los eventos de larga distancia, puesto que no se realizan semifinales). En cada una de estas etapas se hace uso de la función `race_simulation` para simular una carrera, de esta manera obtenemos los que clasifican a la siguiente ronda y nos quedamos con el resultado de la final. Una vez realizadas las n simulaciones de un determinado evento el criterio para ordenar el ranking final es el siguiente: el atleta con mayor cantidad de primeros lugares va primero, en caso de que haya un empate, se pasa a comparar la cantidad de segundos lugares y si persiste el empate, se continúa con los terceros lugares, y así sucesivamente hasta el octavo lugar.

3. Resultados y experimentos

3.1. Hipótesis Extraídas de los Resultados

A partir de los resultados de la simulación, se pueden formular varias hipótesis. Primero, se puede suponer que la recencia de las marcas es un buen predictor del rendimiento actual de los nadadores, ya que los atletas con mejores marcas recientes tienden a obtener mejores resultados. En segundo lugar, la distribución de tiempos parece seguir un patrón identificable y predecible, lo que indica que la KDE refleja correctamente la variabilidad en los tiempos de los nadadores, permitiendo predicciones precisas. Finalmente, la cantidad de marcas registradas por un nadador influye en la confianza del rendimiento, sugiriendo que los nadadores con más marcas tienen predicciones más confiables debido a una muestra de datos más amplia.

3.2. Hallazgo de los resultados

La prediccion de estos juegos olimpicos se torna compleja por varias razones, tenemos la incredible aparicion de jovenes en eventos muy recientes que han hecho marcas muy por encima de los pronosticos (incluso rompiendo records mundiales), los cuales por su extrema juventud cuentan con muy pocas marcas y dificultan la predicción del modelo. Además se suma la reaparición de los actuales campeones olímpicos Caleb Dressel y Emma McKeon los cuales por su escasa cantidad de marcas en las fechas escogidas para la predicción ni siquiera son tomados por el modelo pero no pueden ser ignorados para las cabalas finales.

Cuadro 1. Ranking de los 8 primeros competidores en los 100 metros libres masculinos.

Nombre	1ro	2do	3ro	4to	5to	6to	7mo	8vo
PAN, Zhanle	423	222	114	52	27	22	19	14
POPOVICI, David	285	172	106	73	55	35	33	28
CHALMERS, Kyle	94	116	83	60	66	53	44	36
MIRESSI, Alessandro	45	90	85	76	63	63	44	35
RICHARDS, Matthew	43	71	53	54	46	37	40	60
GROUSSET, Maxime	29	56	37	34	35	46	60	41
ALEXY, Jack	22	32	37	36	31	38	46	62
NEMETH, Nandor	21	65	95	94	68	67	56	27

Cuadro 2. Ranking de los 8 primeros competidores en los 100 metros libres femeninos.

Nombre	1ro	2do	3ro	4to	5to	6to	7mo	8vo
O'CALLAGHAN, Mollie	274	144	80	63	61	34	40	43
HAUGHEY, Siobhan Bernadette	194	104	61	58	59	61	52	32
JACK, Shayna	146	139	103	66	65	52	35	53
MCKEON, Emma	96	121	124	108	94	74	42	57
SJOESTROEM, Sarah	91	140	134	116	88	62	53	14
HARRIS, Meg	58	71	84	86	72	62	42	63
STEENBERGEN, Marrit	39	39	50	46	40	63	54	47
DOUGLASS, Kate	29	42	57	34	39	35	55	48

3.3. Experimentos Realizados para Validar las Hipótesis

Para validar las hipótesis mencionadas, se realizaron varios experimentos. Se compararon los resultados de las simulaciones con competencias recientes, ajustando los parámetros del modelo, como el valor de α, para optimizar la precisión de las predicciones. Además, se utilizó la validación cruzada para seleccionar el mejor bandwidth en KDE, asegurando que el modelo se ajustara adecuadamente a los datos disponibles. También se simularon diferentes escenarios de peso de las marcas, tanto recientes como antiguas, y cantidad de marcas, observando cómo estos factores afectaban los resultados para refinar el modelo.

3.4. Variables de Interés

Las principales variables de interés en el análisis estadístico de la simulación incluyen los tiempos de competencia, convertidos a segundos para facilitar el análisis, y la fecha de la marca, utilizada para ponderar la recencia. También son relevantes el número de marcas registradas por cada atleta, ya que influye en la confianza de la predicción, y los detalles del evento (sexo, distancia, estilo), que afectan las condiciones y resultados esperados.

3.5. Análisis de Parada de la Simulación

El análisis de parada de la simulación es un proceso crítico para asegurar que el número de iteraciones sea suficiente para obtener resultados confiables sin

incurrir en un costo computacional excesivo. En nuestro caso, hemos decidido realizar 1000 iteraciones para cada evento de natación. La validación empírica con datos anteriores respalda esta elección ya que al realizar comparaciones con datosde competencias pasadas los resultados de las simulaciones con dicho número de iteraciones mostraron una buena concordancia con los resultados reales.

4. Modelo Matemático

4.1. Descripción del modelo

Una vez se tienen todas las marcas de un atleta, como se mencionó anteriormente, estas son modificadas en pos de premiar a aquellos que tienen mayor número de marcas, ya que de cierta manera sus marcas son más çonfiables-[1]. Las marcas son modificadas de esta manera:

$$marca_i = marca_i * (1 + \frac{\alpha}{cantMarcas}) \tag{1}$$

.

Para las simulaciones se escogió un valor de $\alpha = 0{,}1$ puesto que al comprobar con competencias anteriores fue el valor que mejores resultados arrojó.

Una vez modificadas las marcas procedemos a utilizar el método no paramétrico de estimación de función de densidad de probabilidad KDE asignándole un peso a cada marca. El peso de cada marca se calcula teniendo en cuenta la fecha en la que se realizó de la siguiente manera:

Primero calculamos los días transcurridos desde la fecha de referencia hasta la fecha dada:

$$d_{fecha} = (fecha - fechaReferencia).days \tag{2}$$

donde:

$$fechaRreferencia = datetime(2022, 1, 1) \tag{3}$$

. Luego calculamos los días totales entre la fecha de referencia y la fecha objetivo:

$$d_{total} = (objetivo - fechaReferencia).days \tag{4}$$

donde:

$$objetivo = datetime(2024, 7, 1) \tag{5}$$

. Normalizamos la cantidad de días de la fecha dada en el intervalo [0, 1]:

$$normalizado = \frac{d_{fecha}}{d_{total}} \tag{6}$$

. Finalmente aplicamos la función logarítmica para obtener el valor:

$$valor = log(normalizado + 1) \tag{7}$$

.

A la hora de escoger el mejor *bandwidth* para aplicar KDE a las marcas de un atleta utilizamos el método *Cross-Validation* y en caso de que el atleta posea pocas marcas utilizamos el método *Leave-One-Out Cross-Validation*. En los casos excepcionales en que un atleta posea una sola marca dejamos un *bandwidth* fijo de 0.2.

Cuando ya tenemos la distribución KDE de las marcas de un atleta en un evento procedemos a tomar muestras de esta para simular las carreras.

4.2. Supuestos

El modelo se basa en varios supuestos clave que permiten simplificar el proceso de simulación: Se asume que el rendimiento de cada atleta en una carrera es independiente del rendimiento de los otros atletas y de sus rendimientos en carreras anteriores. Esto significa que no hay efectos de interacción entre los atletas ni dependencia temporal en sus rendimientos. Se supone que las marcas más recientes son más representativas del rendimiento actual del atleta. Por lo que utilizamos una función logarítmica para ponderar las marcas dándole un mayor peso a las más recientes. Asumimos que la probabilidad de que un atleta haga una marca cercana a sus tiempos reales es uniforme por tanto utilizamos un kernel de tipo *tophat* para el KDE.

4.3. Restricciones

El modelo presenta varias restricciones inherentes a su diseño y al enfoque metodológico: La simulación se repite 1000 veces para obtener una estimación robusta de los resultados. Aunque este número proporciona un buen equilibrio entre precisión y costo computacional, puede ser insuficiente para capturar todas las posibles variaciones en algunas circunstancias. Al dar mayor peso a las marcas más recientes, el modelo asume que el rendimiento reciente de los atletas es más representativo de su estado actual. Teniendo en cuenta que la precisión del modelo depende en gran medida de la cantidad y calidad de los datos disponibles. Cualquier imprecisión o falta de datos puede afectar negativamente la validez de las simulaciones y los resultados obtenidos. Para aquellos casos en que un atleta presente una sola marca se toma un *bandwidth* fijo de 0,2.

Referencias

1. Karla Olivera: Sistema para la Predicción de los Pronósticos del Mundial de Atletismo, pp. 15-16 (2022)