Simulación de competencias de Break Dance

Massiel Paz Otaño
Marlon Díaz Pérez
Albaro Suárez Valdés

8 de junio del 2024

1 Introducción

Este proyecto está encaminado a predecir los resultados de Break Dance en los Juegos Olímpicos Paris 2024. El objetivo es obtener un modelo matemático que describa el comportamiento de cada atleta, basado en los resultados de las competencias en las que ha participado, para poder predecir su actuación futura.
En nuestro proyecto el sistema a modelar representa una comptencia, donde cada atleta constituye una variable que, por el momento, asumimos que son indpendientes entre sí. Además, se analizará si el país en el que se realiza las competencias influye en el resultado (puntos) de los atletas.

2 Detalles de Implementación

2.1 Manejo de datos

Los datos utilizados en este proyecto fueron extraídos de la página oficial de la World DanceSoprt Federation: https://www.worlddancesport.org, los cuales se almacenaron en el archivo DataBboys.csv. En él, se encuentra información relacionada con los atletas y las competencias en las que participó cada uno: *Name* (nombre artístico), *Surename* (nombre real), *Country* (país que representa), *Ranking* (lugar en el que quedó en 1 competencia), *Points* (puntos que obtuvo en dicha competencia), *Date* (fecha en la que se desarrolló la competencia), *Event* (nombre de la comptencia), *Location* (lugar en el que se desarrolló la competencia), entre otros que no serán tomados en cuenta para este primer modelo.
Para recuperar los datos del .csv se utilizó la biblioteca *pandas* de Python.

Con vistas a las simulaciones se crearon 3 métodos que agrupan los datos necesarios de 3 competencias (las 3 más recientes que tuviesen datos): en un diccionario se almacenan como llaves los nombres de los atletas que participaron en la competencia, y como valores los puntos obtenidos en las competencias anteriores a esta. Los puntos se almacenan en orden cronológico, de la competencia más reciente a la menos reciente, de forma que las primeras posiciones en el arreglo corresponden a los puntos obtenidos en las competencias más recientes, y lo contrario para el caso de las últimas posiciones. No existen dos competencias que se hayan realizado en la misma fecha.

2.2 Métodos de simulación

En este proyecto se realizaron 2 métodos de simualaciones bastante similares. Ambos emplean Kernel Distribution Estimation (KDE) para, a partir de los puntos obtenidos en competencias anteriores, obtener una función de distribución que describa su comportamiento para predecir el resultado de la competencia que se esté simulando. Para ello se hizo uso de la biblioteca *KernelDensity* de Python, empleando el 'tophat' como kernel y un bandwidth = 0.75. El segundo método se diferencia del primero en que le otorga un peso mayor (para darle más importancia) a las puntuaciones obtenidas en competencias más cercanas a la que se está simulando. Esto, en teoría, le debiera otorgar un sentido de evolución en el tiempo y fiabilidad a los resultados anteriores, lo que debe resultar en una simulación más precisa. Más adelante se mostrarán los resultados de precisión.
Se han creado, además, 2 funciones (una por cada método de simulación) que permite realizar **n** simulaciones.

2.3 Cálculo de precisión de los métodos de simulación

Primero, realizamos una comparación entre la media de **n** simulaciones y la de los resultados reales, mediante un t-test. Luego, usamos una métrica para estimar la precisión de las simulaciones, basada en probabilidades: se compara la probabilidad de que un atleta haya quedado en el ranking real de la competencia, luego de realizar **n** simulaciones. Para ello, se calcula la probabilidad de que en las **n** simulaciones un atleta A haya quedado en la posición k (posición real en la competencia) dado que no quedó en las *k-1* primeras posiciones, en algunas de las **n** simulaciones. Se asume independencia entre las variables. Además, al resultado de hallar esta probabilidad se le multiplica un parámetro alpha, $0 <= \text{alpha} <= 1$, que dona más peso a los primeros lugares predichos. La suma de todos los alphas es igual (o bastante cercana) a 1.

3 Resultados y Experimentos

3.1 Representaciones visuales

Primeramente, a través de las gráficas de la 1 a la 6 (ir a Anexos) se puede apreciar un comparación visual entre los resultados reales de una competencia y los de realizar 1 simulación. Como se puede ver, el comportamiento de la simulación se asemeja al de los resultados reales, aunque sin ser indénticos (hay puntos en los que difieren en gran medida). Esto nos da la idea de que, si bien una simulación no predice exactamente los resultados reales, tampoco está muy lejos de la realidad.

En segundo lugar, las gráficas de la 7 a la 12 muestran una representación visual de **n** simulaciones de una competencia, efectuadas por un método de simulación. En las primeras 4 puede verse que el margen de puntos otorgados para un mismo atleta es ancho, sin embargo en las últimas 2 (figuras 11 y 12) se puede observar que este margen disminuye, asemejándose un poco más a la gráfica original de puntuaciones obtenidas en la respectiva comptencia (ver gráficas 5 y 6). Esto pudiera sugerir que los resultados de las predicciones para la 3ra competencia son más precisos que para las 2 primeras. (La precisión de cada simulación, por competencia, se podrá apreciar en la siguiente subsección).

3.2 Comprobando la precisión

En un primer momento se realiza un t-test entre la media de puntos otorgados a cada atleta en una competencia por un método de simulación, y las puntuaciones reales, para comprobar si existe, en primera instancia, una diferencia marcada entre las medias de cada array. Se tomaron como hipótesis las siguientes:
H_0: No existe una diferencia significativa entre los resultados originales y los simulados.
H_1: Existe una diferencia significativa entre los resultados originales y los simulados.
Para las 2 primera competencias, este test arroja que para ambas simulaciones no existen datos suficientes para rechazar la hipótesis nula, sin embargo, para la última competencia se tiene que sí se rechaza. Este resultado sugiere que para las 2 primeras competencias, la media de los resultados propuestos por ambos modelos de simulación en **n** simulaciones se asemeja a la media de los resultados reales, por lo que, aparentemente, no hay mucha diferencia, no siendo así para la 3ra competencia en la que, al parecer, los

resultados reales distan de los simulados. Esta métrica no es del todo fiable, ya que al haber un solo valor muy distinto del resto, la media reflejará que la mayoría de los valores se encuentran entre el valor muy distinto y el resto que son ceranos, lo cual no es cierto. Por esta razón se presenta a continuación una métrica más justa.

La siguiente métrica está basada en probabilidades: se calcula la probabilidad de que, en las **n** simulaciones, un atleta *A* haya quedado en la posición *k* (posición real en la competencia) dado que no quedó en las *k-1* posiciones anteriores. Se asume independencia entre las variables (en un primer momento). Al resultado de hallar la probabilidad se le multiplica un parámetro $\alpha, 0 \leq \alpha \leq 1$, que otorga más peso a los primeros lugares predichos. Téngase en cuenta que $\alpha_1 + \alpha_2 + ... + \alpha_n + \epsilon = 1$. Para la elección del valor de α nos apoyamos en la Teoría de números, de modo que $\alpha_i = \frac{1}{p_i}$ donde p_i = i-ésimo número primo.
En la 1ra competencia, esta métrica nos dio para el primer modelo de simulación, una precisión de 0.11 aproximadamente, y de 0.21 (aproximadamente) para el 2do modelo. Con una diferencia del 10%, la cual consideramos estadísticamente significativa, podemos asumir que para esta métrica el segundo modelo es más preciso que el primero. En la 2da competencia, la precisión para el primer modelo fue de 0.31 y, de 0.22, aproxi madamente, para el 2do. Con una diferencia del 9%, que consideraremos estadísticamente significativa, esta métrica sugiere que el primer modelo es más preciso que el 2do. Finalmente, en la 3ra competencia, el 1er modelo tuvo una precisión de 0.19 aproximadamente, mientras que el 2do tuvo un 0.22, luego con una diferencia del 3% que no consideramos estadísticamente significativa, podemos decir que ambos modelos poseen el mismo nivel de precisión.
Resumiendo, a partir de la precisión obtenida para cada modelo en cada competencia, podemos asumir que ambos arrojan predicciones similares; sin embargo, teniendo en cuenta que la 1ra competencia simulada es la más reciente que está registrada en los datos, lo cual implica que hay mayor información para analizar y determinar una distribución aplicando KDE, sumado a que fue la mayor diferencia estadística entre un modelo de simulación y otro, podemos concluir que el modelo que otorga un parámetro de importancia a las puntuaciones de los atletas en dependencia de la fecha en la que la obtuvieron, arroja una predicción más precisa.

3.3 Otros análisis

A modo de perfeccionar la precisión modelos futuros, decidimos realizar un análisis estadístico sobre la influencia del lugar (ciudad-país) en el que se

realizan las competencias sobre el intervalo de puntuación de los atletas que participaron.
Primeramente, en la gráfica 13 se puede apreciar un histograma que registra la cantidad de participantes en una competencia realizada en una ciudad específica. Es posible observar una marcada diferencia entre los números, lo cual puede insinuar que existe un nivel de importancia entre las competencias que puede desencadenar en que los mejores atletas se presentan solo en algunas de ellas. Asimismo se pudiera realizar un análisis sobre cómo varía el desempeño de un atleta en una competencia importante y en otra, no tan importante, y obtener así una medida de cómo ese atleta trabaja en ambientes donde hay más presión. (Dado que el objetivo es predicir el resultado de los juegos olímpicos, esta se considera como una de las competencias más importantes en las que la presión por parte de los atletas está en su máxima expresión).
En segundo lugar, se realizó un test de ANOVA con la biblioteca *scipy.stats* de Python en la que se analiza si hay una diferencia significativa entre los puntos que se otorgaron en competencias realizadas en lugares distintos. Se tomaron como hipótesis las siguientes:
H_0: No hay una diferencia significativa entre los puntos otorgados en una competencias, respecto a los otorgados en otra.
H_1: Hay una diferencia significativa entre los puntos otorgados en una competencias, respecto a los otorgados en otra. Tras arrojar un p-value casi igual a 0, se rechazó la hipótesis nula afirmando que sí existe una diferencia significativa entre las puntuaciones de los atletasa en las 3 ciudades que fueron sedes de 3 competencias. Sin embargo, se hace necesario verificar el cumplimiento de los supuestos para ANOVA para verficar la veracidad de este resultado.

4 Conclusión

Para poder predecir los resultados de una comptencia de Break Dance, se diseñaron dos modelos de simulación basados en KDE, los cuales toman resultados de las competencias anteriores a la simulada en las que participaron los atletas, para predecir su desempeño en la atual. El segundo modelo tiene en cuenta, además, la fecha de las competencias anteriores, otorgando más peso a los resultados de competencias más recientes. Luego de evaluar la preción de ambos en 3 de las competencias más recientes, se llegó a la conclusión de que el 2do aporta resultados más cercanos a los reales. No obstante, en estos primeros modelos se asumió un carácter independiente entre las variables (puntuaciones de cada atleta en una competencia), lo cual no es del

todo cierto, ya que las competencias de Breaking se dividen en batallas donde los bailarines se enfrentan 1 vs 1, y quien gane pasará a la siguiente ronda, luego, el ranking de un atleta dependerá de a quiénes se haya enfrentado en una competencia. Para mejorar la precisión de estos métodos es necesario tenerse en cuenta, además, factores influyentes como el nivel de importancia de la competencia, la ciudad en la que se realiza, así como las "relaciones de competencia" entre los atletas que participan

5 Anexos

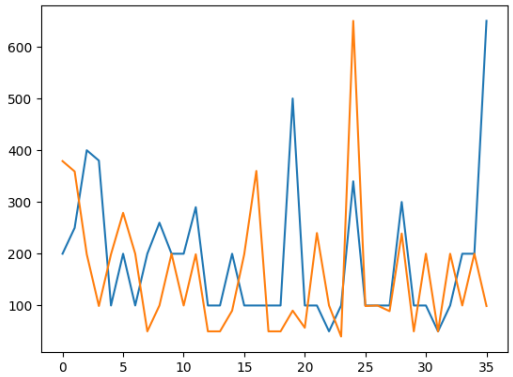

Figure 1: ***BfG World Series del 15-Dec-23***
Simulación sin peso
azul: puntos reales
anaranjado: puntos simulados

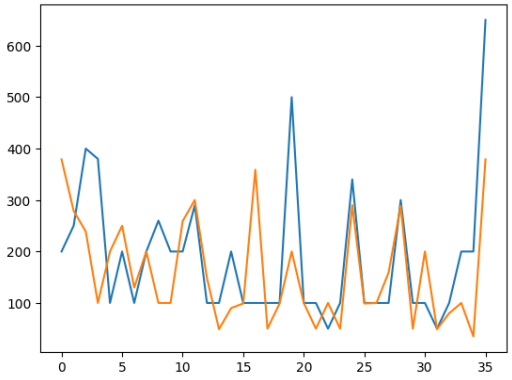

Figure 2: ***BfG World Series del 15-Dec-23***
Simulación con peso
azul: puntos reales
anaranjado: puntos simulados

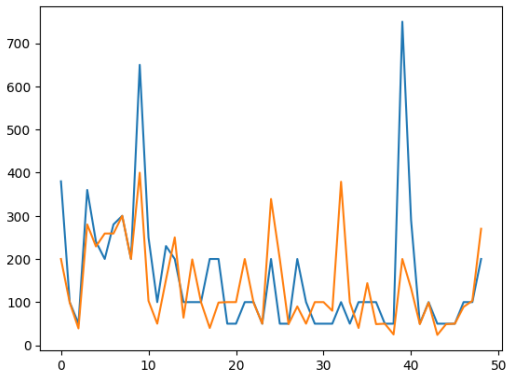

Figure 3: ***BfG World Series del 30-Aug-23***
Simulación sin peso
azul: puntos reales
anaranjado: puntos simulados

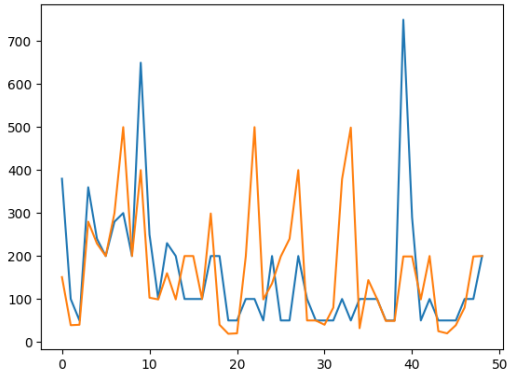

Figure 4: ***BfG World Series del 30-Aug-23***
Simulación con peso
azul: puntos reales
anaranjado: puntos simulados

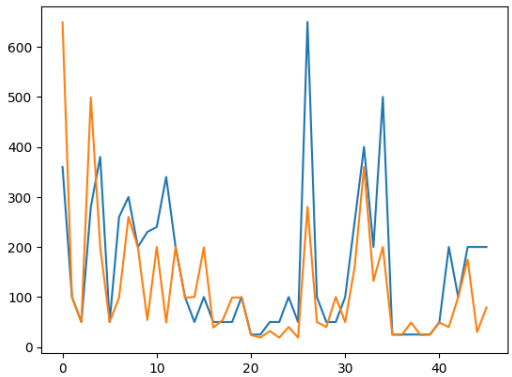

Figure 5: ***Continental Championship del 6-May-23***
Simulación sin peso
azul: puntos reales
anaranjado: puntos simulados

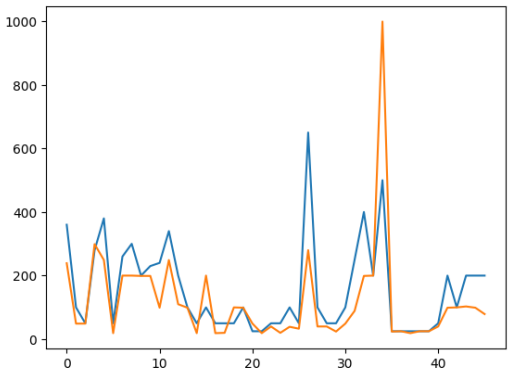

Figure 6: ***Continental Championship del 6-May-23***
Simulación con peso
azul: puntos reales
anaranjado: puntos simulados

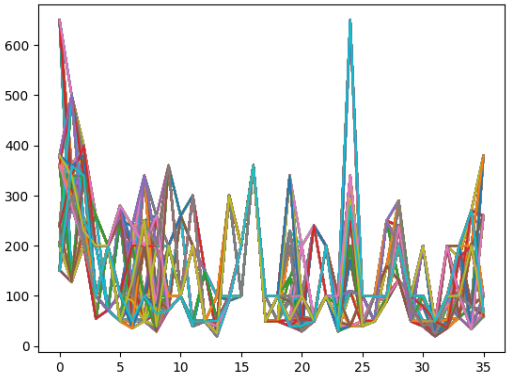

Figure 7: ***BfG World Series del 15-Dec-23***
Simulación sin peso

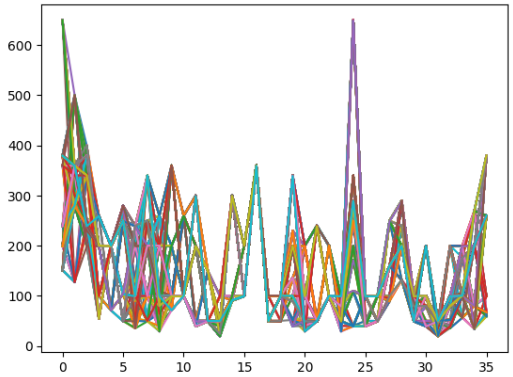

Figure 8: ***BfG World Series del 15-Dec-23***
Simulación con peso

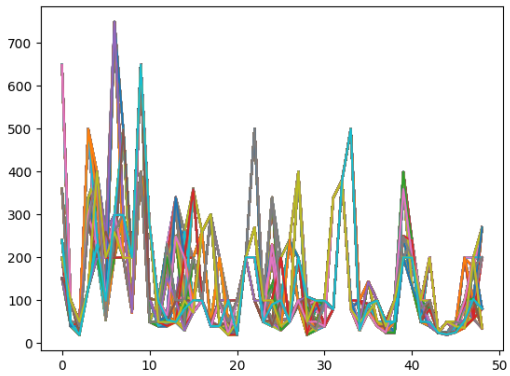

Figure 9: ***BfG World Series del 30-Aug-23***
Simulación sin peso

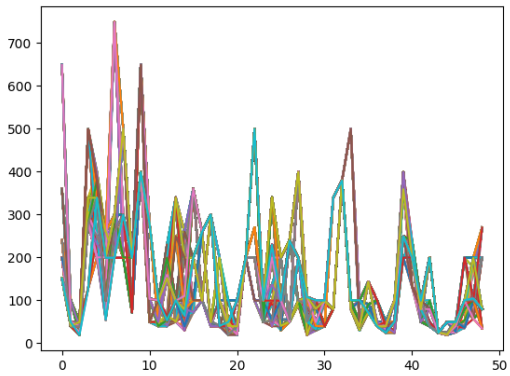

Figure 10: ***BfG World Series del 30-Aug-23***
Simulación con peso

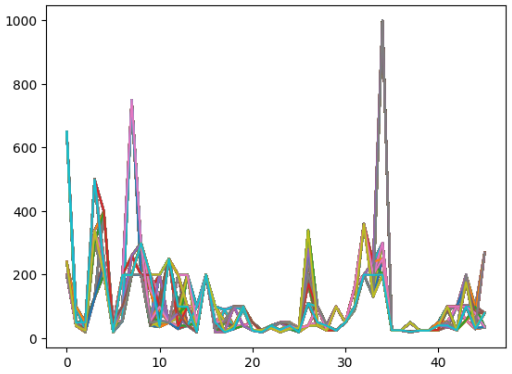

Figure 11: ***Continental Championship del 6-May-23***
Simulación sin peso

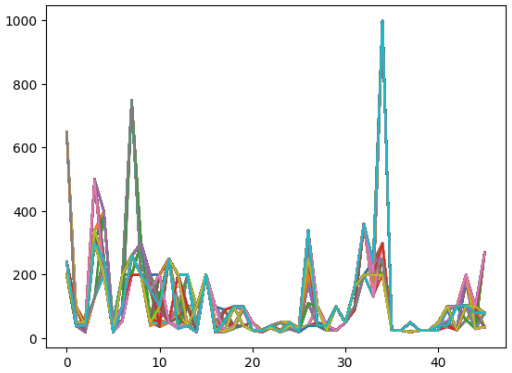

Figure 12: ***Continental Championship del 6-May-23***
Simulación con peso

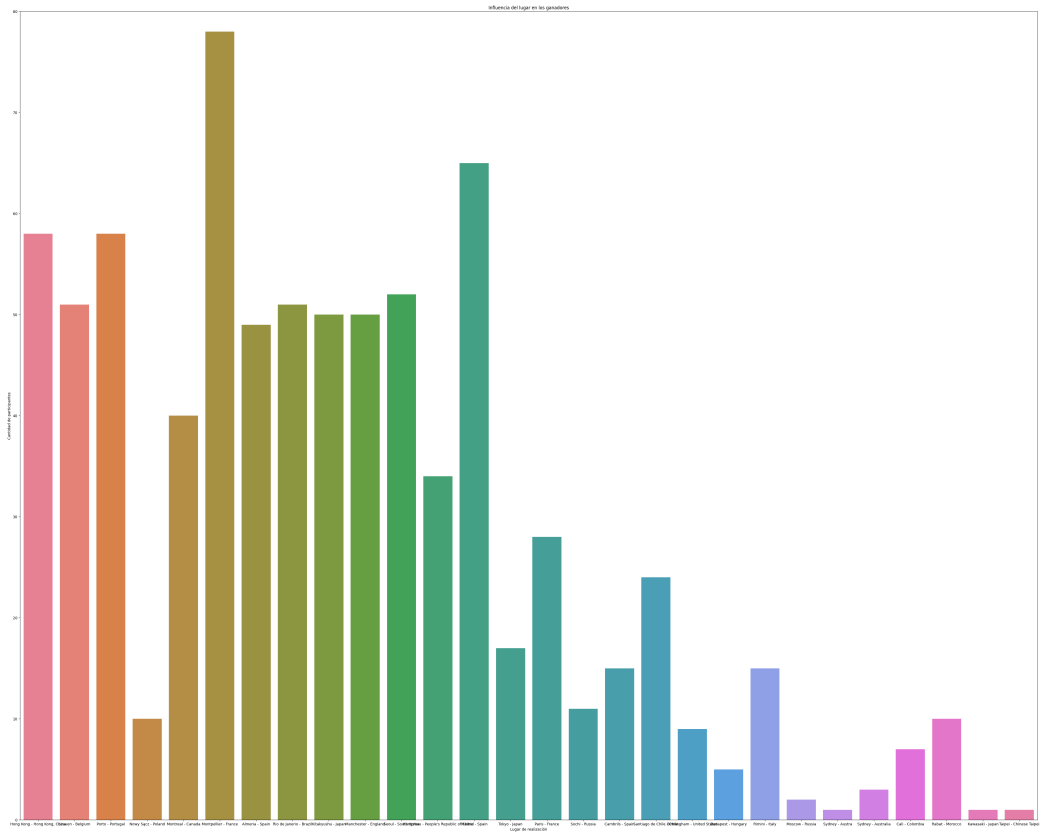

Figure 13: **eje x:** *Ciudades*
eje y: *Cantidad de participantes*