UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

FACULTAD DE CIENCIAS

La ciencia de datos en la predicción de resultados deportivos

PRESENTAN

Vera Santana Juan Manuel
Yañez Hernández Ricardo Jair

5 de junio de 2023

Resumen

En este trabajo se busca el poder predecir un evento deportivo (Mundial Catar 2022) haciendo uso de las herramientas de la Ciencia de Datos, para esto se realizó **web scraping** a las páginas de Wikipedia para poder obtener la información de todos los mundiales jugados, esto con la ayuda de las librerías de **Pandas** y **Selenium**, así como una extensión para Chrome que nos permite abrir las páginas desde la consola y así obtener los datos. También fue necesario conocimiento básico de HTML para obtener los xpath's de la información. Posterior a la obtención de la información, se limpió y usando la **distribución de Poisson**, se generó un modelo matemático que intenta predecir el ganador del Mundial de Catar 2022 pasando por todas las fases de un mundial real. Entre los resultados más importantes está que se pudo predecir al subcampeón del mundial, siendo este Francia

Objetivo General

Usar herramientas de la ciencia de datos para la predicción de resultados en una competencia deportiva, en este caso el mundial de fútbol Qatar 2022.

Hipótesis

El modelo será capaz de predecir la mayoría de los resultados de los partidos del Mundial Catar 2022 e idealmente obtener como campeón a Argentina

Introducción

La Copa Mundial de la FIFA Qatar 2022 fue la vigésimo segunda edición de esta competencia organizada por la FIFA. Tuvo del 20 de noviembre al 18 de diciembre de 2022 en Qatar. Contó con la participación de 32 selecciones distribuidas en 8 grupos (del A hasta el H).

¿Por qué es importante poder predecir quién va a ganar o perder un partido o una competencia? Su importancia radica en el ámbito de las apuestas, por ejemplo. Esto debido a que existen convenios en los que las casas de apuestas (como Caliente.mx, por ejemplo) obtienen derechos de publicidad o transmisión de los partidos, con esto tanto las organizaciones responsables de las competencias (como la Liga MX o en este caso el Mundial de la FIFA) y las casas de apuestas resultan beneficiadas.

Para poder realizar el análisis correspondiente se utilizaron los datos encontrados en Wikipedia mediante web.archive.org sobre la organización de los grupos que posteriormente fueron tratados y organizados mediante la paquetería de Python **Pandas**, puesto que se necesitaron datos de la página en el año 2022. A partir de estos datos se generaron dataframes correspondientes a cada grupo con toda la información necesaria. Posteriormente, mediante **web scraping**, se obtuvo la información necesario con respecto a los partidos pactados para el Mundial de Qatar 2022, es decir, qué selección jugaría contra qué selección. Después se generó un dataframe con la información correspondiente a los mundiales pasados, es decir, desde el mundial del año 1930 hasta el mundial del año 2018 se obtuvieron todos los resultados de los partidos.

Para la elaboración de un modelo de predicción se hizo uso de la **Distribución de Poisson**, que es una distribución discreta que describe el número de eventos que pueden ocurrir en un intervalo de tiempo fijo. En nuestro caso, los eventos son los goles que se anotan dentro de los 90 minutos de duración de un partido de fútbol. La distribución de Poisson es un modelo apropiado siempre y cuando se cumplan las siguientes condiciones:

- la número de veces que ocurre un evento es un número natural.
- que ocurra un evento no afecta que ocurra otro evento, es decir, los eventos son independientes.
- la tasa a la que ocurren los eventos es constante.
- dos eventos no pueden ocurrir en el mismo instante de tiempo

Claramente un partido de fútbol cumple con las condiciones para poder modelar con una distribución de Poisson, puesto que si los eventos corresponden a goles, obviamente en un partido puede haber desde 0 hasta N goles con $N \in \mathbb{N}$. Luego, que un equipo anote un gol no afecta que se pueda anotar otro gol. Después, las probabilidades de que se anoten goles en un partido de 90 minutos son las mismas en todos los partidos. Finalmente, no se pueden anotar dos goles al mismo tiempo.

La distribución de Poisson tiene la siguiente forma

$$P(x) = \frac{\lambda^x e^{-\lambda}}{x!} \tag{1}$$

Donde λ es el promedio de goles en 90 minutos y x el número de goles que podría marcar uno de los dos equipos.

La forma en como se implementó esta distribución es de la siguiente forma: Para la λ esta se obtuvo al multiplicar el promedio de goles de un equipo por el promedio de goles recibidos por su contrincante, y de la misma forma para el otro equipo, el promedio de goles anotados por el promedio de goles recibidos por el otro equipo. Así entre más 'débil' un equipo, más fuerte será su contrincante, y viceversa. Para simular todos los posibles resultados se generó un ciclo for que irá interando todos los posibles resultados de un juego, desde $(0, 10)$ hasta $(10, 0)$ y se irá sumando la probailidad ya sea que gane uno o el otro, o en cuyo caso, haya empate. Se escogió el 10 como el número máximo pues este es la mayor goleada registrada en un Mundial de Fútbol.

Resultados

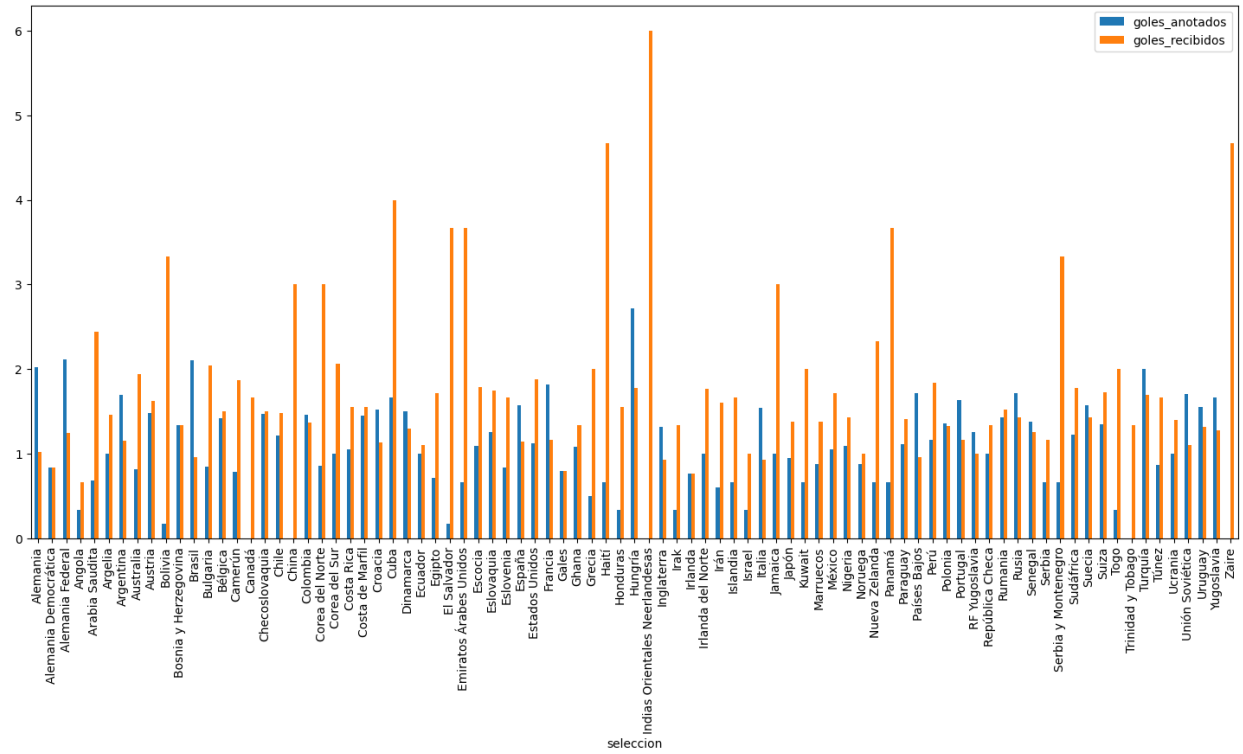

Figura 1: Niveles de poder para cada país que ha participado en el mundial desde su primera edición

En la siguiente tabla se muestran los resultados de la fase de grupos de la simulación:

	Selección	Pts		Selección	Pts
Grupo A	Países Bajos	4	Grupo E	Alemania	7
	Senegal	2		España	5
	Ecuador	2		Japón	3
	Catar	0		Costa Rica	2
	Selección	Pts		Selección	Pts
Grupo B	Inglaterra	6	Grupo F	Croacia	7
	Gales	5		Bélgica	6
	Estados Unidos	3		Marruecos	4
	Irán	2		Canadá	0
	Selección	Pts		Selección	Pts
Grupo C	Argentina	7	Grupo G	Brasil	8
	Polonia	6		Suiza	4
	México	4		Serbia	3
	Arabia Saudita	1		Camerún	2
	Selección	Pts		Selección	Pts
Grupo D	Francia	7	Grupo H	Portugal	6
	Dinamarca	6		Uruguay	5
	Túnez	3		Ghana	4
	Australia	2		Corea del Sur	2

Los resultados para los octavos de final

Selección 1	Selección 2	Avanza
Países Bajos	Gales	Países Bajos
Argentina	Dinamarca	Argentina
Francia	Polonia	Francia
Inglaterra	Senegal	Inglaterra
Alemania	Bélgica	Alemania
Brasil	Uruguay	Brasil
España	Croacia	España
Portugal	Suiza	Portugal

Cuartos de final

Selección 1	Selección 2	Avanza
Alemania	Brasil	Brasil
Países Bajos	Argentina	Argentina
España	Portugal	Portugal
Inglaterra	Francia	Francia

Semifinales

Selección 1	Selección 2	Avanza
Argentina	Brasil	Brasil
Francia	Portugal	Francia

Final (y tercer lugar)

Selección 1	Selección 2	Avanza
Argentina	Portugal	Argentina
Brasil	Francia	Brasil

Análisis y Discusión

Al observar la tabla de puntajes de cada grupo y comparandola con los resultados reales, se obtiene que el modelo predijo correctamente que en el grupo A calificarían los equipos de Países Bajos y Senegal. Para el grupo B el modelo predijo correctamente que Inglaterra calificaría como primer lugar, pero erró que Gales calificaría como segundo, puesto que Estados Unidos calificó como segundo lugar. Para el grupo C el modelo predijo correctamente que Argentina y Polonia calificarían a la siguiente ronda. Para el grupo D, el modelo predijo correctamente que Francia calificaría como líder de grupo, pero erró que Dinamarca calificaría como segundo lugar, puesto que Australia fue quien calificó. Para el grupo E, el modelo erró que Alemania calificaría como primer lugar, puesto que fue Japón quien calificó como primer lugar. Para el grupo F, el modelo erró que Croacia y Bélgica calificarían como primero y segundo respectivamente, puesto que fue Marruecos quien calificó como primer lugar, mientras que Croacia calificó como segundo lugar. Para el grupo G, el modelo acertó que Brasil y Suiza calificarían como primer y segundo lugar respectivamente. Finalmente, para el grupo H, el modelo acertó que Portugal calificaría a la siguiente ronda como primer lugar, pero erró que Uruguay calificaría como segundo lugar, dado que fue Corea del Sur calificó como segundo lugar.

Para los octavos de final, el modelo predijo que Países Bajos clasificaría a cuartos de final, lo cual es acertado. El modelo también predijo acertadamente que Argentina clasificaría a cuartos de final. Del mismo modo, el modelo predijo correctamente Francia e Inglaterra clasificarían a cuartos de final. El modelo erró que España clasificaría a cuartos de final, puesto que fue Marruecos quien clasificó, pero predijo correctamente que Portugal clasificaría.

Para los cuartos de final, el modelo predijo que Brasil y Argentina clasificarían para poder enfrentarse en semifinales, lo cual es errado porque Croacia ganó contra Brasil en cuartos de final, pero predijo correctamente que Argentina clasificaría. Por otro lado, el modelo predijo que Francia y Portugal clasificarían para enfrentarse en semifinales, lo cual es medianamente acertado, ya que Francia sí clasificó a semifinales pero éste jugó contra Marruecos en lugar de Portugal. Para las semifinales, el modelo sólo acertó que Francia clasificaría a la final, puesto que fue Argentina el otro finalista, no Brasil. Finalmente, el modelo predijo que Brasil sería el ganador del torneo, lo cual se aleja de la realidad, puesto que Brasil fue eliminado en cuartos de final y fue Argentina el ganador del torneo.

Adicionalmente, con el modelo se pudo predecir que Argentina quedaría como tercer lugar del torneo. Sin embargo esto es errado porque fue Croacia quien resultó como tercer lugar del torneo.

Conclusiones

En general el modelo trabaja de forma correcta, pues en la fase de grupos logró acertar en un 68.75 % de veces el equipo que avanza a los octavos de final. Tenemos casos como el de Japón que este mundial se presentó como un equipo más fuerte, sorprendiendo a la mayoría del público. Para los octavos logró predecir el partido de Francia vs Polonia, de Inglaterra vs Senegal, Portugal vs Suiza, pero lo que nos importa son los que lograron pasar, en este caso acertó el 75 % de los que pasaron, sólo fallando en el caso de Croacia y de Marruecos. En los cuartos de final logró predecir el 50 % de los partidos jugados, los otros dos no pudo hacerlo debido a que nisiquiera consideró a Croacia y Marruecos. Para las semifinales logró predecir el 50 % de los equipos que pasaron y lamentablemente ninguno de los partidos, esto de nuevo a que no considera a Croacia y Marruecos, aún así no emparejó a Argetina con Francia (cosa que pasó en realidad). En la final ganó Brasil, pero en segundo lugar quedó Francia, como en realidad pasó, en el tercer puesto no lo acertó, porque de nuevo no consideró a Croacia o Marruecos. Pero el top 2 del mundial sigue apareciendo en el top 3 de la simulación, por lo que se puede decir que el modelo trabaja bastante bien

Referencias

[1] Wikipedia. (s.f). *Copa Mundial de Fútbol de 2022*. https://es.wikipedia.org/wiki/Copa_Mundial_de_F%C3%BAtbol_de_2022

[2] FIFA. (s.f). *Llaves y grupos de la Copa Mundial de la FIFA*. https://www.fifa.com/fifaplus/es/tournaments/mens/worldcup/qatar2022/knockout-and-groups