# Modelo predictivo para un partido cerrado de temporada regular 2019-2020 de la NBA

Universidad Nacional de Colombia, Facultad de Ciencias,
Departamento de Física

Carlos Salas, Jaime Cocunubo, Santiago Rodriguez y Jonatan Gonzalez

1 de julio de 2021

**Resumen**

Se plantea como problema para el proyecto realizar un pronóstico acerca del equipo ganador en un partido cerrado de la temporada regular 2019-20 de la NBA a partir de un modelo mecánico estadístico. Para ello se decidió analizar un modelo basado en el movimiento browniano donde se estudiará la diferencia de puntos en un intervalo de tiempo de 48 minutos de juego (duración de un partido sin considerar un posible overtime o tiempo extra) permitiendo estimar la probabilidad de que un equipo gane dado la ventaja de puntos que tuvo en un instante del partido. Se determinó implementar la información de las fluctuaciones en las cuotas de las casas de apuestas con el fin de poder asociar diferentes factores que pueden ocurrir a lo largo del partido. La página Kaggle será utilizada para recolectar los datos necesarios, mientras que para su respectivo tratamiento (limpieza y categorización) se hará uso del entorno de programación Python. Adicionalmente, se precisó implementar el modelo propuesto a 9 partidos del equipo de Los Ángeles Lakers, donde los equipos contrarios corresponden a tres equipos de la misma conferencia: Los Angeles Clippers, Denver Nuggets y Houston Rockets. Por otra parte, se determinó modelar analíticamente la probabilidad según el movimiento browniano utilizando la función de distribución normal, definiendo un parámetro de arrastre según la información de las casas de apuestas y calculando el parámetro correspondiente a la desviación estándar utilizando el método de Newton-Rhapson. Finalmente, se implementará el modelo resultante para después comparar y analizar con los resultados reales.

## 1. Introducción

El baloncesto es uno de los deportes con mayor cantidad de seguidores en el mundo, lo cual hace que las ganancias monetarias en este sean de gran impacto para el mercado. Con esto en mente, se han desarrollado varios modelos con el objetivo de predecir el comportamiento de los equipos durante la temporada y los resultados de cada enfrentamiento con el fin de así lograr un acercamiento al resultado final. Es una de las herramientas más apropiadas, debido al comportamiento aleatorio de las variables, el empleo de la mecánica estadística y los distintos modelos que pueden ser aplicados a partir de la misma. Durante la investigación de los distintos métodos ya aplicados se consiguen tres principales los cuales se adecuan según el objetivo y datos tratados para el abordamiento del problema.

Es usual encontrar en la literatura [1, 2] trabajos que buscan modelar el desarrollo de partidos profesionales mediante distribuciones estadísticas. Sin embargo, estos se suelen ver limitados a la hora de incluir parámetros, que en el caso del baloncesto suelen ser determinantes, entre estos se incluyen lesiones, factores externos como motivación, desgasto físico, agresividad de juego, entre otros. Todos estas variables resultan difíciles de cuantificar e incluir en los modelos, sin embargo, mediante el estudio de las fluctuaciones en las cuotas de las casas de apuestas, es posible cuantificarlos. Adicionalmente, los

modelos usuales tratan de predecir el número total de puntos al final del partido, dado que se utilizan distribuciones que sirven para conteo pero que no diferencian entre eventos para los dos equipos. En ese orden de ideas, se propone utilizar procesos que permitan predecir el resultado final el juego (Gana Local/Visitante) a partir de variables dinámicas que se dan durante los primeros cuartos del juego.

En el primero de ellos se muestra el empleo de diferentes distribuciones de probabilidad con el fin de comprender de una manera global el partido, es decir, se considera exclusivamente el número total de puntos y las variables que determina la dinámica del juego; al hacerlo de esta forma se modela cada cuarto del juego adecuando las distribuciones que mejor se ajustan a los datos recolectados. Los trabajos previos muestran un buen resultado empleando las distribuciones de Poisson, Gamma y Ley de Potencias, sin embargo, no resultan útiles a la hora de predecir el vencedor del partido.

El segundo método está basado en los procesos de Markov, con los cuales se puede analizar el juego en lapsos discretos de tiempo por medio de una matriz estocástica que modela los distintos estados en los que se puede encontrar el sistema y cómo eso afecta la evolución temporal de las variables empleadas para la descripción. Con este método se podría solucionar el problema abierto que ha dejado [8], dado que el artículo analiza el juego de una manera global en cuanto eventos (puntos marcados durante) y por medio de este método es posible predecir cual de los equipos será el ganador.

En el tercer método se aprovecha la naturaleza aleatoria de las variables para el modelo de movimiento browniano o procesos de Weiser con "drift". En ese orden de ideas se relaciona la probabilidad de que el equipo local gane el partido como función de la diferencia entre los puntajes para ambos equipos y el tiempo transcurrido, mediante el uso de la función de distribución acumulada para una distribución normal. Cabe aclarar que para el modelo planteado en [4] el parámetro "drift" se fija como una ventaja para el equipo local, de manera que se propone convertirlo en un factor dinámico relacionado con las cuotas de las casas de apuestas.

Así entonces se plantea como objetivo modelar una serie de partidos a partir de estos métodos con el fin de lograr una buena predicción para los enfrentamientos y su desarrollo en el tiempo, es decir, se busca conocer con mayor grado de certeza los resultados finales del partido de antemano.

## 2. Marco teórico

Para el planteamiento de este proyecto se considerarán *modelos estáticos*, es decir que las variables del sistema no dependerán del intervalo de tiempo de manera local, en pequeños $dt$; el uso de esta clase de modelos reducen las dimensiones y complejidad del sistema a analizar, con la metodología adecuada, se logra emplearlos para sistemas dinámicos como en este caso. Se tiene también varias formas de acercarnos a modelos predictivos en un ámbito deportivo como por ejemplo: el empleo del movimiento browniano [4], un ajuste por medio de distribuciones de probabilidad tales como la Poisson, Beta y Ley de Potencias [2] y modelos mas heurísticos dados por la colectividad. En el presente proyecto se seguirá la metodología de considerar los puntos como una *función de densidad de probabilidad* buscando la *distribución* que de mejor manera se ajuste al sistema en determinados lapsos discretos de tiempo y así predecir la evolución del partido.

Es por este mismo motivo que será de vital importancia mencionar que al tratar las variables del sistema de manera aleatoria y continua estamos conviniendo que la misma es una variable real definida sobre un espacio de probabilidad en el cual se satisface que existe una función real no negativa e integrable $p_x$ tal que para para todo $x$ en los reales se satisface

$$P_x(x) = \int_{-\infty}^{x} p_x(t)dt \tag{1}$$

donde $p_x$ será la función de densidad de la variable aleatoria $x$ [5], la cual por hipótesis estará definida a trozos. Las distribuciones que mejor se ajustan en el desarrollo serán: la Poisson, Beta y Leyes de Potencias de las cuales deberemos tener en cuenta que

- **Distribución Poisson:** En este caso se tendrá en cuenta que la fdd vendrá dada de la forma:

$$p(x) = e^{-\lambda}\frac{\lambda^x}{x!} \tag{2}$$

en caso que $x = 0, 1, 2, ...$, de otra manera $p(x) = 0$. $\lambda$ es el parámetro propio del sistema el cual representa el número de veces que se espera que ocurra el fenómeno durante un intervalo dado de tiempo y es requerido que sea mayor a cero; $x$ es la cantidad de veces de interés que suceda el evento.

- **Distribución beta:** Este modelo es útil ya que logra representar variables físicas cuyos valores se encuentra restringidos a un intervalo de longitud finita, esta distribución se caracteriza por contar con parámetros $a > 0$ y $b > 0$ propios del sistema y con fdd dada de la forma:

$$p(x) = \frac{1}{B_{(a,b)}}x^{a-1}(1-x)^{b-1}\chi_{(0,1)}(x) \tag{3}$$

donde $B_{(a,b)}$ es la función beta

$$B_{(a,b)} = \int_0^1 x^{a-1}(1-x)^{b-1}dx \tag{4}$$

donde se requiere que $0 \leq x \leq 1$, es decir, una normalización de la variable empleada.

- **Distribución Ley de potencias:** Esta es una relación matemática muy simple que permite relacionar variables de la forma

$$p(x) = Cx^p \tag{5}$$

en donde $C$ y $p$ son constantes reales propias del sistema; esta distribución es importante dado que es invariante de escala, es decir, al multiplicar por una constante la variable del sistema, $p(x)$ simplemente se ve afectada por una constante.

- **Distribución Normal:** Para finalizar esta sección de distribuciones se mencionará la distribución normal dada su importancia en el modelo browniano que se trabajará más adelante. Así como con las anteriores se mencionarán sus aspectos más importantes. Se dice que una variable aleatoria $X$ tiene distribución normal de parámetros $\mu$ y $\delta$, donde $\mu$ es un número real y $\delta$ es también un real, pero exclusivamente positivo; su función de densidad (fdd) viene dada por:

$$\Phi(x) = \frac{1}{\delta\sqrt{2\pi}}exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\delta}\right)^2\right], \text{ para x en los reales} \tag{6}$$

en donde se puede identificar a los factores $\mu$ como un factor de localización y a $\delta$ como uno de escala.

Para el análisis de este problema se han empleado distintas herramientas, como ya se mencionaba; una de ellas es el uso de las distribuciones de probabilidad mostradas durante intervalos de tiempo dados de forma periódica, este análisis resulta muy útil cuando el objetivo es analizar de manera global el comportamiento de los eventos, que para este caso sería la cantidad absoluta de puntos, es decir: hechos tanto por el equipo A como el B. Cuando se busca determinar cual de los dos equipos será el ganador, los modelos más útiles serán los de *Markov*[6] y el relacionado al *Movimiento browniano*[4], es por eso que ahora se procederá a dar un marco de referencia al respecto.

El primero de ellos son *Los Procesos de Markov*, al buscar generar un modelo de este estilo se consideran todas las variables necesarias en el desarrollo del partido las cuales puedan determinar el éxito o fracaso de un equipo en la cancha. En particular cuando se habla de un modelo de Markov se hace referencia a un problema en el cual es verificado cuando la probabilidad de un evento se encuentra únicamente relacionada con el evento inmediatamente anterior. En [6] se hace referencia a tres factores que determina el proceso:

1. Posesión del balón, local o invitado.

2. Cómo el equipo ha ganado la posesión del balón.

3. La cantidad de puntos que han sido conseguidos por medio de la anterior posesión.

Y al tener la composición de un estado caracterizado por cada uno de estos 3 factores se conocerá la información básica del partido en ese determinado lapso de tiempo.

Se puede también expresar un proceso de Markov como una serie de experimentos en los que cada uno tiene m posibles resultados y la probabilidad obtenida se encuentra en dependencia exclusiva de los resultados de los experimentos previos. Este proceso se encuentra expresado por una matriz de transición o estocástica, la cual satisface que:

1. $a_{ij} \geq 0$

2. $\sum_j a_{ij} = 1$, para cada i fijo

estos procesos son así conocidos ya que satisfacen la propiedad de Markov, en la cual se establece que en un sistema dinámico dependiente del tiempo las probabilidades de un evento son independientes de sus pasados valores, en otras palabras "las variables aleatorias no tienen memoria", matemáticamente la podemos expresar como:

$$P[X_{n+1} = x_{n+1}|X_0 = x_0, X_1 = x_1, ..., X_n = x_n] = P[X_{n+1} = x_{n+1}|X_n = x_n] \tag{7}$$

Anteriormente se habló de los procesos de Markov y su importancia en la modelación de sistemas dinámicos, y lo cierto es que el modelo del movimiento browniano es un caso particular para procesos de Markov de tiempo y espacio de estados continuos, es decir, procesos de Markov homogéneos en el tiempo.

El movimiento browniano se evidenció por primera vez en el laboratorio del botanista británico Robert Brown, el cual observó que los granos de polen suspendidos en una gota de agua seguían trayectorias aparentemente caóticas. En 1905 Albert Einstein logró explicar analíticamente este fenómeno empírico, sin embargo, la gran mayoría de literatura que se encuentra al respecto habla sobre los procesos del movimiento browniano, conocidos también como de Weiner, el cual en 1923 fundó las bases matemáticas de este proceso estocástico. De modo que para poder aplicarlo al problema de interés será necesario satisfacer ciertas hipótesis. [19]

Por un lado, un proceso estocástico estándar de movimiento browniano debe satisfacer que, dado el proceso aleatorio $\mathbf{X} = \mathbf{X_t} : \mathbf{t} \in [\mathbf{0}, \infty)$:

1. $P(X_0 = 0) = 1$ .

2. $\mathbf{X}$ tiene incrementos estacionarios. Es decir, la distribución del movimiento en un intervalo temporal, depende de la longitud del intervalo.

3. $\mathbf{X}$ tiene incrementos independientes. Es decir, las variables aleatorias son independientes.

4. $X_t$ tiene una distribución normal con media 0 y varianza t para cada $t \in [0, \infty)$.

5. $X_t$ debe ser continua con probabilidad 1.

Es claro que este modelo es útil en muchas situaciones, sin embargo para el caso a estudiar es necesario añadir un factor de arrastre o "drift" $\mu$. Recordando que el movimiento browniano estándar se describe por una distribución normal, este drift se puede interpretar como un desplazamiento en el valor promedio de un proceso aleatorio. Ahora, un cambio evidente para el modelo de movimiento browniano con un drift se puede ver en la condición d para el estándar, de modo que se transforma en [19]:

- $X_t$ tiene una distribución normal con media $\mu t$ y varianza $\sigma^2 t$ para cada $t \in [0, \infty)$.

Donde el parámetro drift puede tomar valores reales $\mu \in \mathbf{R}$ y la varianza $\sigma \in [0, \infty)$. Además, estos parámetros se escogen como funciones lineales de t, dado que ahora $\mathbf{X}$ tiene incrementos independientes y estacionarios del promedio y la varianza.

Otro modo de entender este proceso con drift $\mu$ y varianza $\sigma^2$ fijos, es dando un proceso de movimiento browniano estándar $\mathbf{B} = B_t : t \in [0, \infty)$, de modo que el nuevo proceso con drift será:

$$\mathbf{X}(t) = \mu t + \sigma \mathbf{B}(t) \tag{8}$$

Si ahora se normaliza el tiempo que dura el partido de baloncesto de modo que $t \in (0, 1)$, es posible definir un proceso $X(t)$ de modo que represente la diferencia de puntos entre el equipo local y el visitante, y que por ende que pueda tomar valores negativos, positivos, o el valor nulo [4]. Donde es claro que $X(1) > 0$ indica que el local ganó el partido, y $X(1) < 0$ indica lo contrario. Ahora, si se asume que este proceso se puede describir mediante un modelo de movimiento browniano con drift, se puede asociar el drift por unidad de tiempo $\mu$ a una ventaja o desventaja del equipo local sobre el visitante. Para el modelo presentado por Hal S. Stern (1994), el drift $\mu$ tenía un mismo valor para todos los encuentros estudiados, en el cual se le daba una ventaja arbitraria al local [4]. Dado que en la actualidad se tiene acceso a los datos en tiempo real o "play-by-play data", una propuesta interesante consiste en calcular el parámetro de drift como función de las cuotas en las casas de apuestas para el mercado "Gana Local/Visitante", dado que estas varían dinámicamente como se explicará más adelante.

Luego, continuando con la idea de describir el proceso $X(t)$ como un movimiento browniano con drift, se sabe que se podrá asociar con una distribución normal $X(t) \backsim N(\mu t, \sigma^2 t)$, y de la propiedad **3** de los incrementos independientes, se tendrá que:

$$X(s) - X(t) \backsim N(\mu(s-t), \sigma^2(s-t))$$

Donde $X(s) - X(t)$ para $s > t$ será independiente de $X(t)$. Ahora, recordando que la función de densidad de probabilidad de una distribución normal viene dada por (6), es posible ver que en general para una distribución normal se tendrá la función de distribución acumulada (FDA) $\Phi$ dada como la integral desde $-\infty$ hasta x para la ecuación (6), tal como se procede en la ecuación (1). Obteniendo la FDA normal:

$$\Phi(X, \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left[-\frac{1}{2\sigma^2}(t-\mu)^2\right] dx \tag{9}$$

Luego, definiendo la tasa "drift rate" como el cociente $\mu/\sigma$, la probabilidad de que el equipo local gane el juego vendrá dada como:

$$P(X(1) > 0) = \Phi(\mu/\sigma) \tag{10}$$

Adicionalmente, recordando que el modelo browniano es un proceso de Markov homogéneo en el tiempo, se tendrá que su transición de probabilidad vendrá dada como:

$$p_t(x, y) = f_t(y - x) = \frac{1}{\sqrt{2\pi t}} \exp\left[-\frac{(y-x)^2}{2t}\right] \tag{11}$$

De modo que es extensible para el caso de un proceso de movimiento browniano con drift dado por:

$$f_t(x) = \frac{1}{\sigma\sqrt{2\pi t}} \exp\left[-\frac{1}{2\sigma^2 t}(x - \mu t)^2\right] \tag{12}$$

De forma que es fácilmente demostrado a partir de la condición **4**. Ahora, de la ecuación (10) se puede hallar una equivalencia, ya que para el caso general (un t y un $l(t)$ dados) donde l es la diferencia de puntos entre el equipo local y el visitante, la probabilidad vendrá dada por una probabilidad condicional, es decir, la probabilidad de que local gane, dado que existe cierta diferencia de puntos en un tiempo t. Luego, es posible llegar a la relación:

$$P_{\mu,\sigma} = P(X(1) > 0 | X(t) = l) \tag{13}$$

$$= P(X(1) - X(t) > -l) \tag{14}$$

$$= \Phi\left(\frac{l + (1-t)\mu}{\sqrt{(1-t)\sigma^2}}\right) \tag{15}$$

De manera que cuando $t \to 1$ para $l \neq 0$, entonces la probabilidad debe ser 0 o 1. Ahora, tratando $X(t)$ como una variable aleatoria continua que en cierto modo se aproxima a su valor entero más grande, se arregla el problema respecto de la necesidad de variables continuas. Respecto al modelo presentado por Stern [4], se introduce otro cambio significativo, dado que cuando hay un empate al final del partido, es decir $X(1) = 0$, el modelo considera que las probabilidades de que gane el local son de 0.5. En este modelo se plantea usar una función del drift, la varianza y tendencias anteriores del partido.

Para el problema a abordar se sabe que la variable dependiente puede tomar dos valores binarios, es decir:

$$Si\ X(1) > 0 \Longrightarrow Y = 1$$
$$Si\ X(1) < 0 \Longrightarrow Y = 0$$

Donde $Y = 1$ representa que el local ganó el partido, porque la diferencia de puntos en el tiempo final $t = 1$ fue mayor a cero. Si se tiene en cuenta que el modelo permite encontrar la probabilidad como:

$$P(Y = 1|X) = \Phi(X^T\beta) \tag{16}$$
$$P(Y = 0|X) = 1 - \Phi(X^T\beta) \tag{17}$$

Donde $X^T$ es el traspuesto del vector de variables dependientes, y $\beta$ un vector de coeficientes a ajustar a la regresión.

Finalmente, por medio de la estimación de máxima verosimilitud es posible relacionar la ecuación (13) con las variables transformadas $l/(\sqrt{1-t})$ y $\sqrt{1-t}$ con coeficientes $1/\sigma$ y $\mu/\sigma$. Si se tienen eventos y resultados independientes, la probabilidad vendrá dada como una productoria, de modo que para un conjunto de observaciones se tendrá la verosimilitud:

$$\mathcal{L} = \prod_{i=1}^{n} \left(\Phi(x_i'\beta)y_i)[1 - \Phi(x_i\beta)]^{1-y_i}\right) \tag{18}$$

De modo que tomando los tres primeros cuartos de distintos partido por aparte, se llega a que como son independientes se puede escribir la probabilidad como la productoria:

$$\mathcal{L} = \prod_{i=1}^{n_{partidos}} \prod_{j=1}^{3} \Phi\left(\frac{X_{ij} + (1 - \frac{j}{4})\mu_{ij}}{\sqrt{\left(1 - \frac{j}{4}\right)\sigma^2}})\right)^{Y_i} \times \left(1 - \Phi\left(\frac{X_{ij} + (1 - \frac{j}{4})\mu_{ij}}{\sqrt{\left(1 - \frac{j}{4}\right)\sigma^2}})\right)\right)^{(1-Y_i)} \tag{19}$$

De forma que los parámetros de drift y varianza puedan ser calculados al hacer la derivada total de $\mathcal{L}$ e igualar a cero, con la intención de hallar los máximos globales por los métodos de Newton-Raphson, Bisección, o algún otro que permita hallar ceros de ecuaciones numéricamente. Es claro que para el caso en el que el parámetro de arrastre no lleve índices (es decir, que no sea un parámetro dinámico del partido o el cuarto a considerar) este se puede hallar por el método anterior, sin embargo al hacerlo dinámico se tendría que introducir manualmente en la ecuación (19) y en la del cálculo de la probabilidad.

Con la intención de modificar el parámetro de drift, es posible usar las cuotas de las casas de apuestas, y transformarlas en probabilidades aplicables a los modelos. Para ello existen dos métodos que se utilizan actualmente, el de normalización básica, y el modelo de Shin. Como mostró Erik Štrumbelj en 2014 [9], el modelo de Shin es útil para reducir el error generado por cada casa de apuesta realizando un promedio estadístico entre distintas casas. En sus resultados mostró que para partidos importantes de la NBA, la casa de apuestas online Betfair resultaba ser la más acertada en cuanto a su algoritmo de balanceo de cuotas. Por ende, se proponen usar las cuotas de esta casa, y transformarlas mediante una normalización

básica, de modo que $O = (o_1, o_2, ..., o_n)$ sea un vector con las probabilidades derivadas de las cuotas de la casa de apuestas $\pi_i$. Entonces, para el caso de interés se tienen 3 posibles resultados, victoria Local, victoria Visitante, o Empate, de modo que las cuotas se transforman a probabilidades como $o_i = \frac{1}{\pi_i}$, de este modo, se pueden normalizar las probabilidades como:

$$\beta = \sum_{i=1}^{3} \pi_i \tag{20}$$

De modo que las nuevas probabilidades serán $P_i = \frac{\pi_i}{\beta}$. La razón para usar estas cuotas es que las casas las ajustan (y por ende las probabilidades) para ciertos resultados en vivo, en función de la cantidad de personas que le apuestan, debido a que estas pueden tener información oculta como posibles lesiones de algún jugador, cambios en la moral del equipo, etc.

Finalmente, la precisión de estas predicciones se pueden conocer mediante el puntaje de Brier [9], el cual se define como:

$$(\pi_i, \mathbf{a}) = \frac{1}{n}|(\pi_i - \mathbf{a})|^2 \tag{21}$$

Donde $\pi_i$ representan las probabilidades estimadas, y $\mathbf{a}$ es el vector que contiene el resultado verdadero.

## 3. Estado del Arte

Es bien conocido que desde la antigüedad los deportes han hecho parte de la sociedad en su modalidad de competición. De los primeros registros que se encuentran son unos artefactos para éstos deportes en la antigua China alrededor del 1066 y 771 a.C. Registros también los rastrean hasta Egípto y Persia, es en la antigua Grecia dónde se empieza a consolidar una competición multidisciplinaria donde su ganador sería premiado; esto provoca un gran crecimiento y avance en el juego, dado que la ambición de lograr escalonar en la sociedad y el honor de la victoria se volvieron cardinales para los participantes y todos aquellos que alrededor de la actividad disfrutaban del evento [13]. Siguiendo lo habitual de la sociedad, luego de conseguir lo básico de la supervivencia ésta requiere ocio, el cual o bien era la participación activa de los eventos como competidores, o, como lo era para la gran mayoría, apostar y dejar que el azar aumentase la adrenalina del juego [14]. Con el desarrollo de la probabilidad como una rama aplicada de la matemática respecto al azar, se empiezan a desarrollar modelos para poder conocer o predecir los resultados y poder enmarcar los eventos que se pudieran llegar a dar en el juego, cuantos de ellos eran favorables para poder determinar la certeza de ganar o no [5].

De modelos probabilísticos y su influencia en el análisis de sistemas físicos como los fluidos o termodinámica, nace el primer acercamiento a su aplicación: la teórica cinética. Dicha teoría, bajo grandes supuestos, empieza a obtener resultados acordes a los resultados experimentales basados en las leyes de la termodinámica. Con el tiempo, y acreditado a Ludwig Boltzmann, se logra formalizar la matemática de las leyes aplicando la probabilidad y su rigurosidad hasta entonces forjada [16].

Por toda esta evolución que ha surgido desde la competitividad, el deporte, la matemática y la física se han desarrollado modelos para poder comprender cada aspecto involucrado. Ya que para concretar los parámetros y variables adecuadas que rijan el modelo es necesario tener en cuenta las reglas y formas del juego, se han elaborado distintas investigaciones acerca de los deportes más famosos como el baseball o tenis de mesa [4]. Este trabajo en particular se enfoca en el baloncesto dada la gran cantidad de información que se maneja durante un partido y las estadísticas que diferentes casas de apuestas mantienen en libre acceso para consulta.

Es importante resaltar que modelos de un partido en esta disciplina han sido ampliamente estudiados desde diferentes puntos de vista. En el artículo [3] de 2014 se estudia la evolución de la anotación y los intervalos de anotación en los partidos de NBA correspondientes a las temporadas entre 2005 y 2010. Por otro lado, en el artículo [2] de 2016 se presenta un modelo a partir de una distribución de Poisson para el número de canastas en un intervalo de tiempo, donde se trata la anotación de puntos como un proceso

aleatorio. La tesis doctoral [10] de 2016 expone una red bayesiana para modelar el progreso del total de puntos en un partido de NBA, de forma que determinan la probabilidad de que el total de puntos del partido exceda el establecido por un apostador. Con el mismo enfoque del artículo anterior, en el 2020 se publica el artículo [1], donde se realiza un modelo para la cantidad de puntos en un partido basado en procesos gamma. Un modelo basado en un proceso de Markov se presenta en el artículo [8], donde se busca modelar la progresión de un partido usando la información jugada a jugada de cada equipo. Adicionalmente, en el artículo [4] se estudia la diferencia de puntos en un partido a través de un modelo basado en el movimiento browniano.

Aunque como se expuso anteriormente, un partido de baloncesto como objeto de estudio ha sido trabajado en diferentes investigaciones, también existen modelos de partidos deportivos competitivos diferentes que contribuyen al desarrollo del mismo problema. En el 2017 se publica el artículo [11], donde se presenta un modelo de conteo para las anotaciones de un partido de football usando la distribución de Weibull. En el artículo [8], mencionado anteriormente, se expone la posibilidad de aplicar el modelo basado en el movimiento browniano a otros deportes, en particular se aplica a un partido de baseball. Por otra parte, la tesis [12] de 2014 realiza un modelo de un partido de baseball a partir de cadenas de Markov para partidos de la MLB (Major League of Baseball) de 2013.

Esto ha permitido el estudio y análisis de posibles deficiencias en los modelos o nuevas propuestas para procurar un resultado más acertado. Con esta información como punto de partida se puede concluir que un modelo mecánico-estadístico más acertado según los resultados experimentales sería el modelo basado en el movimiento browniano, tal como se propone en el artículo [4] donde se aborda el problema de predecir el resultado de un partido de NBA en temporada regular.

## 4. Planteamiento del problema

El proyecto propuesto plantea el estudio de un partido de baloncesto de NBA (National Basketball Association) como sistema físico compuesto de dos sistemas interactuantes que corresponden a cada equipo. Un análisis de un partido de NBA a partir del proceso gamma se presenta en la referencia [1], donde se trata el problema de modelar el proceso de la cantidad total de puntos al final del partido. Por otro lado, la referencia [2] expone un estudio del número de canastas en un intervalo de tiempo, donde la mayoría de canastas siguen una distribución de Poisson pero en el último minuto de partidos cerrados son distribuidas siguiendo una Ley de Potencias. Considerando el primer artículo se resalta que el modelo permite dar un pronóstico de la cantidad de puntos durante un partido sin diferenciar que equipo anota cada punto. También se destaca que el modelo del segundo artículo considera cualquier tipo de anotación (1 punto, 2 puntos o 3 puntos) como una canasta en un intervalo de tiempo.

Se presenta como problema la realización de un modelo mecánico-estadístico que permita pronosticar el equipo ganador durante un partido de NBA que cumpla la condición de un partido cerrado implementando la información de las fluctuaciones en las cuotas de las casas de apuestas, en contraste con el artículo [1] que modela la cantidad total de puntos del partido y el artículo [2] que estudia el número de canastas en un intervalo del partido. Como primera tentativa se considera el modelo de Poisson y de Ley de Potencias, sin embargo, este no tiene en cuenta el tipo de canasta ni el equipo que la anota, por lo que no se propone como un modelo efectivo para la realización del problema planteado. La referencia [1] plantea la posibilidad de aplicar el proceso gamma a los procesos de anotación local como visitante de forma separada, no obstante, en busca de abarcar otros modelos para el desarrollo del problema se utilizará un modelo basado en un proceso de movimiento browniano.

Se modelará la diferencia entre el puntaje de cada equipo en un partido de baloncesto como una variable aleatoria $X(t)$ en un intervalo de tiempo transformado a unidad $t \in (0, 1)$, donde se tratará $X(t)$ como una variable continua. Asumiendo que se puede modelar la $X(t)$ como un proceso de movimiento de browniano, se puede tomar $X(t)$ como una distribución normal de forma que la probabilidad de que el equipo local gane $P(X(1) > 0)$ estará dada por la función de distribución correspondiente, es decir, $P(X(1) > 0) = \Phi(\mu/\sigma)$. Específicamente, usando el modelo dado por la caminata aleatoria del movimiento browniano, la probabilidad de que el equipo local gane dado que hay una diferencia de puntos $m$

al haberse jugado una fracción del partido está dada por [4]:

$$P_{\mu,\sigma}(m,t) = \Phi\left(\frac{m+(1-t)\mu}{\sqrt{(1-t)\sigma^2}}\right) \tag{22}$$

La probabilidad de que el equipo contrario gane se obtendrá calculando el complemento de la probabilidad ya encontrada.

El parámetro drift o de arrastre en este modelo permite tener en cuenta muchos factores que se presentan de manera especial en un partido, así se propone escoger un parámetro dependiente del tiempo teniendo en cuenta las fluctuaciones de las cuotas en las casas de apuesta, pues factores como la lesión de un jugador o los problemas de faltas se verán reflejados en las apuestas en tiempo real. La forma en la que se hará la conversión de la información de las casas de apuestas a el parámetro $\mu$ del modelo será inicialmente por tanteo, y para luego determinar que tan bueno fue el $\mu$ escogido se comparará el resultado de aplicar dicho modelo con el propuesto en el artículo [4] que toma un parámetro de arrastre independiente del tiempo.

La ecuación (22) se puede interpretar como un modelo probit relacionando el resultado del juego con las variables transformadas $m/\sqrt{1-t}$ y $\sqrt{1-t}$ tal como es propuesto en el artículo [4], donde se asume que las observaciones generadas para cada cuarto del partido son independientes. Con este análisis se espera poder calcular el parámetro $\sigma$ de la distribución haciendo una estimación por máxima verosimilitud, es decir, calculando los máximos de la función de verosimilitud $L$ correspondiente. Para calcular los ceros de la derivada de la función de verosimilitud se utilizará el método de Newton-Rhapson con un programa en Mathematica.

El ajuste que se plantea para el parámetro $\sigma$ se implementará a 7 partidos cerrados de un equipo determinado de la NBA para la temporada de 2019-2020 (considerados partidos de entrenamiento para el modelo), donde los equipos contrarios corresponden a tres distintos equipos de la misma conferencia que terminaron la temporada regular 2018-2019 en una posición similar. El modelo será aplicado a los siguientes 2 partidos cerrados restantes que se jugaron entre estos equipos durante la temporada. Por otro lado, la recolección de datos se hará usando la página Kaggle que permite obtener toda la información requerida de los partidos jugada a jugada de la temporada, mientras que la información de las casas de apuesta se recolectará a través de la página nowgoal3. Para el manejo, la limpieza y la categorización de los mismos se utilizará el lenguaje de programación Python y la librería Pandas. Por último, los pronósticos de cada modelo se compararán con los resultados reales de cada partido de forma que se pueda hacer un análisis de que tan acertado fue cada modelo y cómo se comparan entre sí para así determinar que tan fiable es el modelo.

## 5. Motivación y justificación

El creciente acceso a bases de datos confiables y fáciles de manejar, se ha visto reflejado en el interés de la comunidad científica por crear modelos que describan el comportamiento de diversos sistemas mediante ajustes basados en estos datos. De acuerdo a Mordor Intelligence (2021) [20], el mercado del análisis deportivo fue valorado en 2015 por 83.6 millones USD, y creció en el 2020 a un valor de 1.05 billones USD. Siendo un mercado liderado por empresas como IBM, SAP SE, y Oracle Corporation, la implementación de modelos predictivos en los deportes se ha visto escalada mayormente a sistemas complejos adaptativos, impactando sistemas naturales (el sistema inmune, ecosistemas, sociedades) y sistemas artificiales (inteligencia artificial, redes neuronales, sistemas de computo distribuidos y paralelos)[17, 18].

La motivación para incluir un parámetro de arrastre dinámico al trabajo realizado por Hal S. Stern (1994)[4], surgió del trabajo de F.S Abril y C. J. Quimbay [7], donde se introdujo un parámetro de arrastre estocástico para una serie de tiempo no estacionaria, de modo que en este caso se variase temporalmente

la media de la distribución normal. En ese orden de ideas, al implementar el parámetro de arrastre mediante la fluctuación de las cuotas en las casas de apuestas, y al entrenar el modelo con datos de partidos anteriores, se espera obtener resultados distintos para cada equipo. Con esta solución se espera poder extender el modelo a otros sistemas y áreas, donde también es posible cuantificar de algún modo variables complejas mediante ciertos indicadores o tendencias.

Un ejemplo para cuantificar y clasificar variables que pueden ser complejas recaen en el método One Hot Encoding, el cual crea una columna para cada valor distinto que exista en la característica que se está codificando, si se toma una fila específica tendrá 0 en las columnas que no cumplan la condición y 1 en el caso contrario. Un ejemplo de lo anterior podría ser codificar el sexo de las personas, donde para un hombre se le asigna un 1 a la columna "hombre" y un 0 a la columna "mujer" y lo contrario cuando se trate de una mujer [15].

## 6. Objetivo general

Pronosticar el resultado de una partido regular y cerrado de la temporada 2019-2020 de la NBA, empleando un modelo mecánico-estadístico basado en el movimiento browniano que tenga en cuenta la información suministrada por la fluctuación de las cuotas en las casas de apuesta y la diferencia de puntos en un dado momento del partido.

## 7. Objetivos específicos

1. Recolectar los datos de los partidos de la temporada 2019-2020 de la NBA usando la página Kaggle.
2. Limpiar y categorizar los datos obtenidos de acuerdo a los 4 equipos elegidos: Lakers, Nuggets, Rockets y Clippers, empleando el entorno Python y la librería Pandas.
3. Normalizar el tiempo del juego de cada partido.
4. Escoger 7 partidos de entrenamiento y los demás partidos (2) para probar el modelo desarrollado.
5. Modelar analíticamente la probabilidad de que gane el equipo local en un partido de baloncesto de acuerdo al movimiento browniano, tomando como partida la distribución normal y máxima verosimilitud.
6. Definir el parámetro de arrastre $\mu$ dependiente del tiempo a partir de la probabilidad brindada por las casas de apuestas.
7. Implementar computacionalmente las distribuciones y funciones desarrolladas de manera analítica para calcular la diferencia de puntos de todos los partidos elegidos.
8. Determinar el parámetro correspondiente a la desviación estándar mediante el método computacional de Newton-Raphson con los partidos de entrenamiento.
9. Ejecutar el modelo con los partidos de prueba y dar un pronóstico de equipo ganador.
10. Comparar y analizar las estimaciones realizadas con los resultados reales para determinar la fiabilidad del modelo.

## 8. Metodología

1. Para la obtención de los datos se hace uso de la página Kaggle, la cual es una comunidad de ciencia de datos con una amplia variedad de áreas de conocimiento, en dicha página, específicamente en el link: `https://www.kaggle.com/schmadam97/nba-playbyplay-data-20182019`, se realiza la descarga de los datos correspondientes a la temporada 2019-2020 en formato de extensión csv.

2. Una vez cargados los datos se procede a usar la librería Pandas, la cual permite agrupar los datos en estructuras de dos dimensiones llamadas DataFrames donde se proceden a eliminar los valores repetidos y los partidos de post-temporada (playoffs) mediante funciones predeterminadas con el fin de limpiar la información a usar. Posteriormente, para categorizar, se seleccionan los partidos donde jugaron los Lakers (LAL), Clippers (LAC), Nuggets (DEN) y Rockets (HOU) entre sí, donde se obtuvo un total de 11 partidos en donde los Lakers fueron locales 5 veces y visitantes 6 veces.

3. Dado a que en los datos descargados cada cuarto de juego empieza en 720 segundos y termina en 0, se procede a realizar la conversión temporal donde se tiene que todo el partido comprende 2880 segundos y se busca ordenar el tiempo de manera continua, tal que el primer cuarto empiece en el segundo 0 y el partido finalice en el segundo 2880 para posteriormente dividirlo entre 2880 con el fin de que el intervalo de tiempo esté entre 0 y 1.

4. Para la elección de los partidos de prueba, se visualizan las fechas de los 9 partidos tomados y se toman los últimos dos partidos de acuerdo a la fecha cuando se desarrollaron, teniendo que estos partidos corresponden al 30 de julio y al 10 de agosto del año 2020.

5. En primer lugar se considera la diferencia de puntos entre el equipo local y el equipo visitante en función del tiempo como un proceso estocástico estándar del movimiento browniano. Luego, introduciendo un "drift" o factor de arrastre en el modelo, se llega a que la probabilidad de que un equipo gane un partido como local (dada por la distribución normal), es dependiente del tiempo, la diferencia de puntos $l(t)$, y el arrastre $\mu(t)$.

   Teniendo en cuenta que la variable dependiente es de tipo binario (local gana o local pierde), y que la probabilidad de que local gane dados ciertos parámetros corresponde a la función de distribución acumulada normal. Es posible usar la regresión probit para ajustar la desviación estándar al modelo a partir de los resultados conocidos para los 7 partidos de entrenamiento. Así, se propone hallar la desviación estándar mediante la estimación de máxima verosimilitud.

6. Inicialmente para hallar la probabilidad de las casas de apuestas, se toman de la página `http://www.nowgoal3.com/basketball/1x2-362685` las tasas de apuesta para cada partido en específico en distintos instantes de tiempo. La normalización del inverso de dichas tasas corresponderá a la probabilidad brindada por las casas de apuestas, con dicha probabilidad se define el parámetro $\mu$ mediante la ecuación (8), haciendo que el parámetro $\mu$ sea guardado en una matriz y dependa del tiempo y partido a usar debido a que la probabilidad de las casas de apuestas varía con el tiempo y el partido.

7. Antes de crear la función que permita calcular la probabilidad de que gane el equipo local se mapean los puntos de cada equipo en cada partido y se calcula la diferencia de puntos dependiente del tiempo y del partido, donde se propone almacenar dicha información en una matríz para después llamar elemento a elemento de esta.
   Con lo anterior realizado y el parámetro de arrastre definido, se define una función en Python que dependa de $\mu$, $\sigma$, la diferencia de puntos de juego, el tiempo y tenga la forma de la ecuación (22), lo cual corresponde a la función de distribución acumulativa de una distribución normal estándar que es bien conocida y encontrada en algunas de las referencias de este proyecto. Dicha función será de vital importancia ya que permitirá calcular la probabilidad de que gane el equipo local.

8. Al tener las diferencias de puntos $X_{ij}$, el equipo ganador $Y_i$ y el parámetro de arrastre $\mu$ para los partidos de entrenamiento, se procede a aplicar el método computacional de Newton-Raphson, en donde a partir de un error, máximo de iteraciones e intervalo a trabajar definidos se halla la raíz de la ecuación de interés, determinando entonces la desviación estándar o parámetro $\sigma$.

9. Una vez con todas las variables y parámetros determinados se aplica la ecuación (22) con los datos de los equipos de prueba para obtener una gráfica de probabilidad en función del tiempo y observar cual es la probabilidad de que gane el equipo local en cada partido cuando este ha finalizado, es decir cuando t=1.

10. Una vez obtenidos los pronósticos del equipo ganador para cada partido de prueba, se procede a ver que tan alta o no es dicha probabilidad en cada caso y ver que tanto varía con respecto al resultado real para así decidir si las predicciones realizadas fueron aceptables o no.

Nota: Además del notebook adjunto, se puede encontrar el código de Python en: `https://github.com/sarodriguezme/Proyecto-M.Estadistica` específicamente el archivo *Proyecto_NBA.ipynb*, es posible que en algunas ocasiones no cargue a la primera vez pero volviendo a cargar la página debería aparecer. También se puede utilizar el siguiente código QR:

## 9. Resultados Esperados

Con el desarrollo del proyecto se esperan lograr los siguientes resultados:

1. Con la obtención de datos de la página kaggle se obtienen conocimientos sobre una comunidad de ciencia de datos, donde a demás de poder descargar archivos también es posible crear una cuenta, interactuar con usuarios y los proyectos que estos realicen tanto para complementar información y conceptos como para ayudar en caso de que ese usuario lo pida o presente errores en sus desarrollos. Adicionalmente hay convocatorias para participar en torneos, los cuales consisten en dejar tratar un problema abierto con unos datos en específico.

2. Para el desarrollo del proyecto es imprescindible usar un entorno que permita el manejo de datos, donde se escogió el entorno Python y la librería Pandas. Manejando dicho entorno se busca comprender el concepto de DataFrame y como a partir de funciones y comandos propios de Pandas es posible visualizar, limpiar y manejar a gusto los datos necesarios para el proyecto.

3. Cuando se normaliza el tiempo se busca entender que es un proceso necesario para la correcta ejecución de los métodos del proyecto, principalmente dado a la naturaleza de la función de distribución normal y los parámetros que caracterizan esta.

4. Entender el concepto de sobre ajuste, si se entrena un modelo con unos datos determinados y se prueba solamente en dichos datos, el modelo va a acostumbrarse a estos datos, hará buenas predicciones pero será ineficaz a la hora de usarlo con datos nuevos. Por esto se busca escoger partidos de entrenamiento y prueba en lugar de usar todos los partidos.

5. Con el planteamiento analítico se espera generalizar el modelo añadiendo un parámetro dinámico de arrastre. Además, al implementar la regresión probit y el método de mayor verosimilitud se aprende una técnica que puede ser usada en campos como Machine Learning y Redes Neuronales.

6. Al implementar el parámetro de arrastre $\mu$ como función del tiempo y de cada juego mediante las cuotas en las casas de apuestas, se espera poder realizar un mapeo efectivo, de forma que las predicciones del modelo se ajusten mejor a los resultados reales. Parte de ese trabajo se debe hacer por tanteo o ensayo y error (casi como introducir una función de proporcionalidad experimentalmente), cosa que históricamente se ha hecho muchas veces en la física.

7. El hecho de recorrer un archivo para calcular la diferencia de puntos en distintos tiempos para varios partidos como también implementar un función computacional dependiente de ciertos parámetros que haga el mismo papel de una distribución normal representa un reto que requiere destrezas en el uso de librerías numéricas como Numpy y en aplicar la correcta sintaxis para que el programa reproduzca lo que se busca.

8. Se espera obtener una desviación estándar estática que se ajuste a los partidos de entrenamiento. Un posible valor esperado es $sigma = 15.82$, obtenido por [4] para 493 partidos de la NBA en el año 1992. Este objetivo permite también adquirir conocimientos en la aplicación de métodos numéricos para hallar raíces de ecuaciones no lineales.

9. De la descripción para la dinámica de la probabilidad de tener cierta diferencia de puntos entre local y visitante $P_l(t)$, se espera poder ver una correlación entre la probabilidad y el valor para el parámetro de arrastre, de forma que haya una corrección al modelo más simple. En el desarrollo de esta sección se fortalecerán habilidades de visualización de resultados.

10. De ser posible, se espera comparar la discrepancia en los resultados del método sin arrastre y los del método con arrastre, respecto a los resultados reales de algunos partidos. Se espera que la discrepancia para los resultados del método sin arrastre sea menor, demostrando una mejora con el nuevo método. Acá se espera aprender herramientas para medir la efectividad de un modelo a un problema con resultados binarios (También muy útil en problemas de clasificación en Machine Learning y regresión logística).

## 10. Cronograma

| Objetivo a cumplir | Tiempo estimado |
|---|---|
| Recolectar los datos de los partidos de la temporada 2019-20 de la NBA usando la página Kaggle. | Realizado |
| Limpiar y categorizar los datos obtenidos de acuerdo a los 4 equipos elegidos: Lakers, Nuggets, Rockets y Clippers, empleando el entorno Python y la librería Pandas. | Realizado |
| Normalizar el tiempo del juego de cada partido. | Realizado |
| Escoger 7 partidos de entrenamiento y los demás partidos (2) para probar el modelo desarrollado. | 03/07/2021 |
| Modelar analíticamente la probabilidad de que gane el equipo local en un partido de baloncesto de acuerdo al movimiento browniano, basado en la distribución normal y máxima verosimilitud. | 12/07/2021 |
| Definir el parámetro de arrastre $\mu$ dependiente del tiempo a partir de la probabilidad de las casas de apuestas. | 16/07/2021 |
| Implementar computacionalmente las distribuciones y funciones desarrolladas analíticamente para calcular la diferencia de puntos de todos los partidos elegidos. | 20/07/2021 |
| Determinar el parámetro correspondiente a la desviación estándar mediante el método computacional de Newton Raphson con los partidos de entrenamiento. | 23/07/2021 |
| Ejecutar el modelo con los partidos de prueba y dar un pronóstico de equipo ganador. | 25/07/2021 |
| Comparar y analizar las estimaciones realizadas con los resultados reales para determinar la fiabilidad de cada modelo. | 30/07/2021 |

## 11. Recursos disponibles

- La página Kaggle, `https://www.kaggle.com/schmadam97/nba-playbyplay-data-20182019` donde se obtienen los datos de los partidos correspondientes a la temporada 2019-2020, pero si el lector lo requiere puede descargar datos desde la temporada del año 2015 hasta el año 2020.

- Computadores y el lenguaje de programación Python para procesar datos en entornos de trabajo como Google Colab y notebooks de Jupyter, para este último se usó la aplicación Anaconda Navigator (anaconda 3).

- Librerias de Python como numpy (`https://numpy.org/doc/stable/reference/index.html`) y pandas (`https://pandas.pydata.org/docs/reference/index.html#api`) fueron utilizadas así como los links adjuntos con el fin de facilitar, procesar y tratar la información de una manera práctica y precisa.

- La página nowgoal3, `http://www.nowgoal3.com/basketball/1x2-362685` donde se obtienen los datos de las tasas de las casas de apuestas dependientes del tiempo para los partidos a estudiar.
- GitHub, el cual es una página web donde es posible crear repositorios propios y hacer más sencillo controlar versiones, trabajar en grupo para un proyecto y poder compartirlo con más personas para facilitar su visualización sin necesidad de tener que descargar ningún código o archivo, se puede visualizar incluso desde el celular, puede probar con el código QR que está al final de la sección de metodología.
- Adicionalmente se cuenta como recurso disponible las referencias citadas.

**Referencias**


[1] K. Song, Y. Gao, J. Shi, *Making real-time preditions for NBA basketball games by combinig the historical data and bookmarker's betting line*, Physica A, 2020.
1, 8

[2] J. Martín-González, Y. deSaá Guerra, J. García-Manso, E. Arriaza, T. Valverde-Estévez, *The Poisson model limits in NBA basketball: Complexity in team sports*, Physica A, 2016.
1, 2, 7, 8

[3] Y. de Sá Guerra, J. M. Martín González, S. Sarmiento Montesdeoca, D. Rodriguez Ruiz, N. Arjonilla-López, J. M. García-Manso, *Basketball scoring in NBA games: an example of complexity*, Journal of Systems Science and Complexity, 2013
7

[4] Hal S. Stern, *A Brownian motion for the progress of sports scores*, Journal of the American Statistical Association, 1994.
2, 3, 5, 6, 7, 8, 9, 13

[5] L. Blanco, *Probabilidad*, Universidad Nacional de Colombia, 2004.
2, 7

[6] K. Shirley, *A Markov Model for Basketball*, Columbia University, Applied Statistics center.
3

[7] Abril, F. S. and Quimbay, C. J., *Temporal fluctuation scaling in nonstationary time series using the path integral formalism*, American Physical Society, 2021
9

[8] Petar Vracar, Erik Strumbelj, Igor Kononenko, *Modeling basketball play-by-play data*, Expert Systems With Applications, 2016.
2, 8

[9] E. Strumbel, *On determinig probability forecast from betting ods*, Internal Journal of Forecasting, 2014.
6, 7

[10] E. M. Alameda, *A dynamic Bayesian network to predict the total points scored in national basketball association games*, Iowa State University, 2019.
8

[11] G. Boshnakov, T. Kharrat, I. McHale, *A bivariate Weibull count model for forecastting association football scores*, International Journal of Forecasting, 2017
8

[12] D. Ursin, *A Markov model with applications*, University of Wisconsin-Milwaukee, 2014.
8

[13] Wikipedia, *Deportes*, 2021, `https://es.wikipedia.org/wiki/Deporte#cite_note-8` 7

[14] Red História, *¿Cuál es el origen de las casas de apuestas?*, 2020, https://redhistoria.com/cual-es-el-origen-de-las-casas-de-apuestas/ 7

[15] Interactive Chaos, n.d, *One Hot Encoding*, https://interactivechaos.com/es/manual/tutorial-dne-learning/one-hot-encoding
10

[16] Kubo. Ryogo (1965), *Statistical Mechanics*, Springer. 7

[17] M. Oldham A. T. Crooks, *Drafting agent-based modeling into basketball analytics*, George Mason University. Department of Computational and Data Sciences.
9

[18] M. Wright, *OR analysis of Sporting Rules - A Survey.*, European Journal of Operational Research 232(1): 1-8.
9

[19] Random, K. Siegrist, *Brownian Motion with Drift*, https://randomservices.org/random/brown/Drift.html
4

[20] Mordor Intelligence, *SPORTS ANALYTICS MARKET - GROWTH, TRENDS, COVID-19 IMPACT, AND FORECASTS (2021 - 2026)*, https://mordorintelligence.com/industry-reports/sports-analytics-market

9