# Modelos matemáticos para la predicción de resultados deportivos: Aplicación en distintos torneos de básquet

Tesis de licenciatura en matemática aplicada

**Alejandro Alvarez**

Director: Dr. Guillermo Duran

Buenos Aires 2024

**Indice**

# 1 Introducción

El básquet es un deporte muy popular en Argentina y en el mundo. La liga estadounidense, más conocida como NBA, es el torneo más importante de este deporte.
Los partidos de esta disciplina consisten en dos equipos de 5 jugadores que compiten entre sí por obtener más puntos en un tiempo fijo. Para conseguir estos puntos se debe meter una pelota en un aro ubicado dentro de un área en campo rival a 3.05 metros de altura. Esto otorgará distinta cantidad de puntos dependiendo del tipo de tiro realizado:

- **Triple:** vale 3 puntos, es el tiro desde afuera del área contraria.
- **Doble:** vale 2 puntos, es el tiro desde dentro del área rival.
- **Tiro libre** vale 1 punto, es el tiro realizado después de una falta, se lanza desde la línea de tiros libres.

La duración de los partidos es de 40 minutos, 48 para los de la NBA, dividido en 4 períodos (cuartos), donde cada equipo tiene un 24 segundos para intentar anotar puntos. Al comienzo del partido, para definir quien obtiene la primer posesión se realiza un salto, es decir, el arbitro arroja la pelota hacia arriba y dos jugadores, uno de cada equipo, saltan por conseguirlo. En los 3 cuartos posteriores la primer posesión se alterna es decir en el tercero la tendrá el equipo que ganó el salto, pero en el segundo y en el cuarto la tendrá el contrario.

En este deporte se suelen tomar muchas estadísticas que resultan muy útiles para describir los partidos o a los equipos:

- **Puntos:** Cantidad de puntos anotados por un jugador o por un equipo durante el juego.
- **Asistencias:** Número de pases que terminan en punto.
- **Efectividad de tiros:** Porcentaje de los tiros realizados que logró convertir un jugador o un equipo en un partido, se suelen separar en triples, tiros libres y tiros de campo (junta los tiros de 2 y 3 puntos).
- **Rebotes:** Un rebote se consigue cuando un jugador agarra la pelota luego de que alguien falló su tiro al aro. Se divide en ofensivos, cuando el tiro lo realizó un jugador de tu equipo y defensivos cuando se consigue luego del fallo del rival.
- **Robos:** Un robo se consigue cuando un jugador le quita la pelota al equipo rival, directamente de las manos o cortando un pase.

- **Tapones:** Un tapón se consigue cuando el jugador bloquea el tiro al aro del rival, consiguiendo que no ingrese en el aro.

Gracias al avance de la tecnología y los datos que se toman, en el último tiempo los modelos matemáticos aplicados a deportes tomaron mayores relevancias en equipos profesionales para el armado de las alineaciones de los equipos o prevenir lesiones de los jugadores o para la predicción del resultado de los partidos a jugarse.

En este trabajo, nos centraremos en este último tipo de modelos matemáticos y su uso aplicado en los partidos del Mundial de básquet 2019, el Super20 2019, las temporada 2019-2020 y 2020-2021 de la Liga Nacional y la temporada 2021-2022 de la NBA. Está organizado de la siguiente manera: en la **Sección 2** se hace un paneo breve sobre algunas herramientas estadísticas muy utilizadas utilizadas para esta problemática. En la **Sección 3** se presentan algunos trabajos realizados en esta área de estudio terminando con el algoritmo desarrollado el cual le otorga una probabilidad de victoria a los equipos que se enfrentan en un partido en base a la fecha que se juegue, la localía y los partidos anteriores. En la **Sección 4** se muestran los resultados conseguidos al testear nuestro modelo en las distintas competencias, mundial de 2019, varias copas y ligas de la Liga Nacional de Básquet y por último la comparación con el modelo desarrollado por los matemáticos Kai Song, Qingrong Zou y Jian Shi en 2018 en predicciones de partidos de la NBA y el torneo nacional. Luego en la **Sección 5** se muestra el interés que se generó cuando pronosticamos los partidos del mundial y las ligas nacionales en la página de internet 280777, lugar donde aprovechamos para divulgar nuestro trabajo y el potencial de las matemáticas. Y por último en la **Sección 6** daremos las conclusiones y los posibles pasos a seguir posteriores a este trabajo.

# 2 Marco teórico

En esta sección vamos a dar un breve repaso a algunos conceptos que serán necesarios para comprender mejor los modelos matemáticos usados para la predicción de resultados de los partidos de básquet.

## 2.1 Distribución de Poisson

Una distribución de Poisson modela el número de veces que ocurre un evento en un intervalo de tiempo.
Sea $X \sim Poisson(\lambda)$ entonces su función de probabilidad es $P[X = k] = \frac{e^{\lambda}\lambda^{k}}{k!}$
donde $k = 0, 1, 2, \ldots$ es el número de ocurrencias del evento o fenómeno.

El parámetro $\lambda > 0$ representa el número de veces que se espera que ocurra el fenómeno durante el intervalo dado. Por ejemplo, si el suceso estudiado tiene lugar en promedio 4 veces por minuto y estamos interesados en la probabilidad de que ocurra $k$ veces dentro de un intervalo de 10 minutos, usaremos un modelo de distribución de Poisson con $\lambda$ = 10x4 = 40.

## 2.2 Regresión Lineal

La regresión lineal es un método estadístico que se emplea para modelar la relación lineal entre una variable dependiente $Y$ (también conocida como la respuesta) y una o más variables independientes $X_i \quad i = 1 \ldots n$ (denominadas predictores). El objetivo principal es encontrar la ecuación de una línea recta que mejor se ajuste a los datos observados. Puede ayudar a comprender y predecir el comportamiento de sistemas complejos o a analizar datos experimentales, financieros y biológicos.
La ecuación general correspondiente a un modelo de regresión lineal es:

$$Y = \beta_0 + \sum_{i=1}^{n} \beta_i X_i + \epsilon_i$$

donde $\beta$ representa los parámetros lineales que se deben calcular y $\epsilon$ representa los términos de error.

Los modelos de regresión lineal se basan en los siguientes supuestos:
1. Los errores se distribuyen normalmente.
2. La varianza es constante.

## 2.3 Regresión Lineal Generalizada (GLM)

Este método es una generalización flexible de la regresión lineal ordinaria que permite variables de respuesta que tienen modelos de distribución de errores distintos de una distribución normal. El GLM generaliza la regresión lineal al permitir que el modelo lineal esté relacionado con la variable de respuesta a través de una función de enlace y al permitir que la magnitud de la varianza de cada medición sea una función de su valor predicho.
Estas técnicas se utilizan como una forma de unificar otros modelos estadísticos, como la regresión lineal, la regresión logística y la regresión de Poisson.

### Modelo Poisson

Entre los datos con estructura no normal se encuentran los conteos de sucesos que se producen por azar con cierta frecuencia y son modelizables en términos de tasas de incidencia que dependen de ciertas variables predictoras. Para modelar este tipo de datos se utiliza la distribución de Poisson, $X \sim Po(\lambda)$ donde $\lambda$ representa el número medio de ocurrencias, de forma que $E(X) = \lambda$ y $V(X) = \lambda$

Para éste tipo de datos se puede utilizar un modelo lineal generalizado para variables aleatorias Poisson o más conocido como regresión de Poisson. Sea $Y \sim Po(\lambda)$ una variable aleatoria y sea $X_1, X_2, \ldots, X_p$ un conjunto de variables predictoras que pueden afectar el comportamiento de $Y$ asumimos que:

$$\lambda = exp(\beta_0 + \beta_1 X_1 + \cdots + \beta_2 X_2 + \cdots + \beta_p X_p)$$

donde $\beta_i$ indica el efecto de la variable predictora.
En esta situación y tomando como función de enlace el logaritmo tenemos que:

$$log(E(Y)) = log(\lambda) = \beta_0 + \beta_1 X_1 + \cdots + \beta_2 X_2 + \cdots + \beta_p X_p$$

donde se ve la expresión del modelo lineal generalizado.

## 2.4 Simulación

La Simulación es la imitación del funcionamiento de un sistema real durante un intervalo de tiempo. Esta técnica puede realizarse ya sea de forma manual o computacional.
El comportamiento de la simulación está determinado por el modelo planteado o conjunto de supuestos concernientes al sistema real, estos supuestos se expresan a través de relaciones lógicas y matemáticas entre las entidades.
Tradicionalmente, el modelado formal de sistemas se genera a través de un modelo matemático, que intenta encontrar soluciones analíticas a problemas que permiten la predicción del comportamiento de un sistema con un conjunto de parámetros y condiciones iniciales. La simulación por computadora se emplea con frecuencia en sistemas de modelado que carecen de soluciones analíticas simples y cerradas, dado que la enumeración exhaustiva de todos los estados posibles en el modelo resultaría altamente compleja. En este contexto, se recurre a la generación de una muestra de escenarios representativos mediante simulación computacional, permitiendo abordar de manera eficaz problemáticas para las cuales las soluciones analíticas convencionales no son viables.
En esta tesis se usará para estudiar y predecir los comportamientos de las distintas competencias a predecir.

# 3 Modelos de predicción de resultados deportivos

En esta sección se mostrarán algunos de los trabajos ya realizados por otros investigadores en el área de predicción de partidos de básquet y fútbol, para terminar describiendo en detalle el modelo de Poisson utilizado para nuestros cálculos, empezando por el paper de "Dixon-Coles", pasando por el utilizado en 301060 para llegar a la adaptación obtenida mediante los distintos análisis realizados por nosotros.

## 3.1 Estado del arte

Para la predicción de los partidos de básquet se vienen desarrollando varios trabajos, existen modelos simples basándose en regresiones logísticas hasta modelos de alta complejidad llegando a redes neuronales.

Algunos ejemplos son [KS06] que utilizan una regresión logística y cadenas de Markov para este cometido o también se encuentra [SZS18] que utiliza el supuesto de que el número de goles anotados por el equipo local y visitante en un juego particular puede describirse como un vector aleatorio normal bivariado.
En el trabajo de Horvat, T. [HJLL23] se propone un modelo de aprendizaje automático para predecir resultados de la NBA basado en un índice extendido de eficiencia del equipo. Este índice considera el rendimiento de los jugadores y comparaciones asimétricas con el equipo rival y en [LBB22] los autores realizan las predicciones utilizando diversos tipos de redes neuronales, como retroalimentación directa, de base radial, probabilísticas y de regresión generalizada y también exploran la fusión de redes neuronales mediante redes de creencias de Bayes y redes neuronales probabilísticas.

Por último, ya alejándonos de las predicciones de los resultados, hay varios trabajos que buscan predecir las lesiones de los jugadores en los partidos, como el caso de [CSF21] el cual aborda el desafío de muestras pequeñas al predecir lesiones de jugadores de la NBA con el modelo de aprendizaje profundo utilizando datos longitudinales de lesiones pasadas, actividad de juego y estadísticas de los jugadores.

## 3.2 Modelo de la normal multivariada para predicción en básquet

Investigando algunos trabajos orientados específicamente a la predicción de partidos de básquet nos interesamos en el paper escrito por Kai Song, Qingrong Zou y Jian Shi en 2018 [SZS18].
Para generar su modelo, se usa el supuesto de que el puntaje del equipo local y visitante en un juego particular puede describirse como un vector aleatorio normal bivariado.
Este supuesto de normalidad puede verse extraño sabiendo que toma valores en el continuo de $\mathbb{R}$ y los puntajes son números naturales, pero analizando en detalle el uso y los parámetros de estas

distribuciones podemos notar que la probabilidad de que el resultado sea negativo es extremadamente baja.
Con los datos utilizados en este trabajo, el equipo con la mayor probabilidad de obtener un puntaje negativo para este modelo es Bahía Basket jugando como visitante con $\mu = 75.97$ y $\sigma = 8.08$ quedando una probabilidad $P(Y < 0) = 2.61 * 10^{-21}$.
Luego el hecho de que los puntajes no sean enteros, puede salvarse considerando que la probabilidad de que el equipo $i$ obtenga $n$ puntos es igual a $P(n < Y_i < n + 1)$, pero en este trabajo no será necesario debido a que sólo se considera la probabilidad de ganar o perder el partido.

Los parámetros para estas distribuciones son determinadas por las siguientes 5 estadísticas de rendimiento:
**Porcentaje efectivo de tiros de campo;** $eFG\% = \frac{FG+\frac{1}{2}3P}{FGA}$
**Porcentaje de Tiros Libres;** $FT\% = \frac{FT}{FTA}$
**Porcentaje de pérdidas;** $TOV\% = \frac{TOV}{FGA+TOV+0.44FTA}$
**Porcentaje de rebotes ofensivos;** $ORB\% = \frac{ORB}{ORB+DRB_o}$
**Porcentaje de rebotes defensivos;** $ORB\% = \frac{ORB}{DRB+ORB_o}$

Las estadísticas base que se usan para calcular las anteriores son: tiros de campo convertidos (FG), tiros de campos intentados (FGA), triples convertidos(3P), pérdidas (TOV), rebotes ofensivos (ORB) y rebotes defensivos (DRB). El prefijo $o$ denota al equipo contrario.

El modelo presentado en el paper [SZS18] supone que los puntos convertidos por el local y por el visitante en un partido siguen una distribución normal bivariada ($N_2$) donde las medias dependen linealmente de las estadísticas de rendimiento principales y la matriz de covarianzas está determinada por un parámetro de correlación por localía, común para todos los encuentros, como así también por las varianzas de los participantes según su rol.

Formalmente, sean $i$ y $j$ los índices del equipo local (1) y visitante (2) que juegan en un partido $k$. Sean

- $Y_{1,i}^{(k)}$ y $Y_{2,j}^{(k)}$ los puntos convertidos por cada equipo
- $\mu_{1,i}^{(k)}$ y $\mu_{2,j}^{(k)}$ las medias
- $\Sigma_{i,j}^{(k)}$ la matriz de covarianzas
- $\sigma_{H,i}$ y $\sigma_{A,j}$ las varianzas
- $X_{H,i}^{(k)} = (1, x_{H,1}^{(k)}, \cdots, x_{H,5}^{(k)})^T$ y $X_{A,j}^{(k)} = (1, x_{A,1}^{(k)}, \cdots, x_{A,5}^{(k)})^T$ sus estadísticas de rendimiento.
- $\rho$ es el coeficiente de correlación entre locales y visitantes
- $\beta_H$ y $\beta_A$ son los coeficientes desconocidos correspondientes al local y visitante.

Luego el modelo es de la forma:

$$\begin{cases} (Y_{1,i}^{(k)}, Y_{2,j}^{(k)}) \sim N_2\left(\begin{bmatrix}\mu_{1,i}^{(k)} \\ \mu_{2,j}^{(k)}\end{bmatrix} \quad \Sigma_{i,j}^{(k)}\right), \\ \mu_{1,i}^{(k)} = X_{H,i}^{(k)T}\beta_H, \\ \mu_{2,j}^{(k)} = X_{A,j}^{(k)T}\beta_A, \\ \Sigma_{i,j}^{(k)} = \begin{bmatrix}\sigma_{H,i}^2 & \rho\sigma_{H,i}\sigma_{A,j} \\ \rho\sigma_{H,i}\sigma_{A,j} & \sigma_{A,j}^2\end{bmatrix}, \end{cases}$$

La probabilidad de que el equipo local $i$ le gane al equipo $j$ en el partido $k$ puede calcularse cómo:

$$\begin{aligned} P\left(Y_{1i}^{(k)} - Y_{2j}^{(k)} > 0\right) &= 1 - P\left(Y_{1i}^{(k)} - Y_{2j}^{(k)} \leq 0\right) \\ &= 1 - P\left(\frac{Y_{1i}^{(k)} - Y_{2j}^{(k)} - (\mu_{1i}^{(k)} - \mu_{2j}^{(k)})}{\delta} \leq \frac{-(\mu_{1i}^{(k)} - \mu_{2j}^{(k)})}{\delta}\right) \\ &= 1 - \Phi\left(\frac{-(\mu_{1i}^{(k)} - \mu_{2j}^{(k)})}{\delta}\right) \end{aligned}$$

Siendo $\Phi$ la función de distribución acumulada de una normal estándar y $\delta$ la desviación estándar de $Y_{1i}^{(k)} - Y_{2j}^{(k)}$ que es igual a $(\sigma_{H,i}^2 + \sigma_{A,j}^2 - 2\sigma_{H,i}\sigma_{A,j})^{\frac{1}{2}}$

Una limitación del modelo es que estas estadísticas de rendimiento son desconocidas antes de que se jueguen los partidos, pero pueden estimarse por ejemplo mediante una regresión lineal o con la media de los datos históricos de los equipos, distinguiendo si jugaron de local o de visitante.

## 3.3 Modelo implementado en fútbol: Dixon-Coles

Por la creciente popularidad de las apuestas en el fútbol inglés, se buscaron distintas formas de poder predecir los resultados exactos de los partidos.
Dos de los tantos matemáticos que se propusieron conseguir un modelo capaz de predecir estos resultados, o mejor dicho, las distintas probabilidades de que ocurra un resultado en un partido en particular, fueron Mark Dixon y Stuart Coles.

Lo que ellos propusieron en su paper [DC97] es un modelo matemático en el que se usa el estimador de máxima verosimilitud de los parámetros de dos Poisson independientes que representan la distribución de los goles del local y del visitante en un partido. La media está modelada como una función de los resultados previos de los respectivos equipos. También este modelo tiene en cuenta el efecto de localía, según las estadísticas, los equipos suelen tener mejores actuaciones

cuando juegan en sus estadios y para imitar éste fenómeno se agrega una variable global que se adiciona al equipo local.

Sean las variables aleatorias $X_{i,j}$ y $Y_{i,j}$, los puntos que meten el equipo local y visitante respectivamente y se definen;

$$X_{i,j} \sim Poisson(\alpha_i \beta_j \gamma)$$

$$Y_{i,j} \sim Poisson(\alpha_j \beta_i \gamma)$$

las cuales se asumirán independientes y donde $\alpha_i, \beta_j > 0 \; \forall i$ donde $\alpha_i$ se denominará a la fuerza de ataque de de cada equipo, $\beta_i$ será la de defensa del equipo y $\gamma$ el parámetro que explica el factor de la localía.

Cómo se asume la independencia entre las variables, Maher en su paper de 1982 propone el uso de una familia de Poisson bivariadas como una extensión del modelo básico, pero luego de contrastar el modelo con los datos, se ve que para juegos de baja puntuación la independencia no se corresponde directamente, en 0-0, 1-1 (sub-estiman), 1-0 y 0-1 (sobre-estiman). Una causa de esto se debe a que un equipo que va perdiendo 1 a 0 al final de un partido va arriesgar más que si estuviera 0 a 0, y esta familia es incapaz de representar estos casos.

Por lo tanto en Dixon and Coles se propone la siguiente modificación del modelo;

$$Pr(X_{i,j} = x, Y_{i,j} = y) = \tau_{\lambda,\mu}(x,y) \frac{\lambda^x exp(-\lambda)}{x!} \frac{\mu^y exp(-\mu)}{y!}$$

Donde $\lambda = \alpha_i \beta_j \gamma$, $\mu = \alpha_j \beta_i$ y

$$\tau_{\lambda,\mu}(x,y) = \begin{cases} 1 - \lambda\mu\rho & si \quad x = y = 0, \\ 1 + \lambda\rho & si \quad x = 0 \quad y \quad y = 1, \\ 1 + \mu\rho & si \quad x = 1 \quad y \quad y = 0, \\ 1 - \rho & si \quad x = y = 1, \\ 1 & si \;\; no. \end{cases}$$

Donde $\rho$ es el parámetro de dependencia, tal que,

$$max(-1/\lambda, \;\; -1/\mu) \leq \rho \leq min(1/\lambda\mu, \;\; 1)$$

Notar que si $\rho = 0$ indica que los goles de un equipo son independientes a los de su rival.
Es fácil comprobar que las distribuciones marginales correspondientes siguen siendo Poisson con medias $\lambda$ y $\mu$ respectivamente. Los parámetros de ataque y defensa de cada equipo se consideran constantes en el tiempo que dura el partido.

Una limitación estructural del modelo son los parámetros estáticos, es decir, los equipos tienen un rendimiento constante para todos los partidos después de calcular los parámetros, cuando se conoce que en realidad es dinámico, tienen buenas y malas rachas. Este comportamiento se incorpora en el modelo, recalculando las potencias de los equipos a medida que se va jugando el torneo, por ejemplo, unos días antes de que el partido a predecir se juegue.
Además debe tenerse en cuenta que los partidos recientes brindan mejor información sobre el nivel de los equipos que los más antiguos, para arreglar esto de una forma relativamente sencilla tomaremos que los parámetros son localmente constantes a lo largo del tiempo pero que los datos históricos tiene menos valor que los recientes.

Por lo tanto, a partir de nuestra base de datos, conseguiremos los parámetros deseados usando la siguiente función de "pseudo-probabilidad" (pseudolikelihood);

$$L_t(\alpha_i, \beta_i, \rho, \gamma; i = 1, \ldots, n) = \prod_{k \epsilon A_t} \{\tau_{\lambda_k, \mu_k}(x_k, y_k) exp(-\lambda_k) \lambda_k^{x_k} exp(-\mu_k) \mu_k^{y_k}\}^{\phi(t-t_k)}$$

Donde $t_k$ es el tiempo en el que el partido $k$ fue jugado, $A_t = \{k : t_k < t\}$, $\lambda_k = \alpha_{i(k)}\beta_{j(k)}\gamma$ y $\mu_k = \alpha_{j(k)}\beta_{i(k)}$ y $\phi$ es una función decreciente en el tiempo, la cual se buscará de la forma $\phi(t) = exp(-\xi t)$.
Luego si $\xi = 0$ el modelo es estático y para valores mas grandes de $\xi$ se le da menos peso a los resultados antiguos. Se puede demostrar que maximizando el valor de este parámetro se obtiene $\xi = 0,0065$

### 3.4 301060

El proyecto 301060 es un desarrollo de investigadores y tesistas del Instituto del Cálculo (UBA-CONICET), que se encuentra en la Facultad de Ciencias Exactas y Naturales de la Universidad de Buenos Aires. La idea original surgió del sitio 851.cl, un proyecto con el mismo espíritu, llevado adelante por científicos de la Universidad de Chile y del Instituto Sistemas Complejos de Ingeniería (ISCI) de Chile.

El sitio consiste en poner en marcha modelos matemáticos, inspirados en el paper de Dixon y Coles, que tomen como insumo la historia reciente de los resultados obtenidos por los equipos que participaron del Mundial Qatar 2022 para estimar las probabilidades de los distintos resultados de cada partido. Dados esos cálculos, se simuló el torneo millones de veces para aproximar probabilidades de los posibles resultados parciales o finales.

## Modificaciones en el modelo de predicción

En el modelo del paper antes explicado, los autores consideran que jugar de local implica un beneficio, el cual se suma a la fuerza de ataque del local y defensa del visitante en el parámetro de la distribución del equipo que juega en cancha propia, pero no distingue entre los distintos competidores.

Por esta razón para el nuevo modelo, los integrantes del Instituto del Cálculo propusieron cambiar esta forma de ver la localía por una donde se separa un plus de ataque y un plus de defensa, los cuales no necesariamente son un beneficio, hay equipos que obtienen mejores resultados cuando juegan de visitantes.

Luego de estas modificaciones las probabilidades de un partido de fútbol quedan definidas con la distribución de Poisson bivariada, tal que la probabilidad de que un partido el equipo local $i$ meta $x$ goles y reciba $y$ contra el equipo $j$ es

$$P(X = x, Y = y) = \frac{(\lambda^x * e^{-\lambda})}{x!} * \frac{(\mu^y * e^{-\mu})}{y!}$$

Donde $\lambda = exp(\alpha_i - \beta_j + \gamma_i)$ y $\mu = exp(\alpha_j - \beta_i - \delta_i)$.

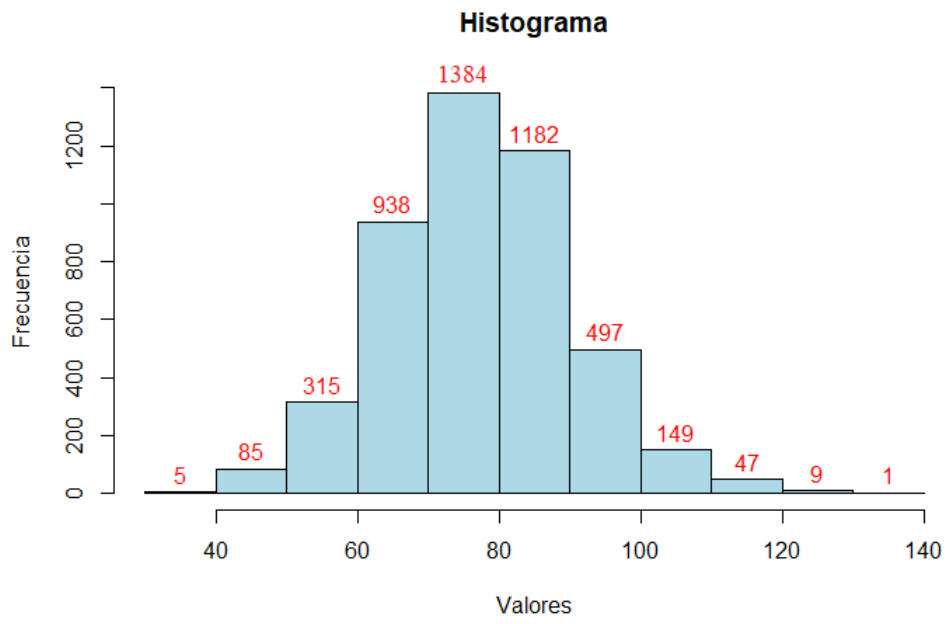
Siendo $\alpha_i$ es el poder de ataque del equipo $i$, $\beta_i$ es el poder de defensa del equipo $i$ (que afecta al equipo rival), $\gamma_i$ el plus de ataque del equipo local $i$ y $\delta_i$ el plus de defensa del equipo local $i$.

### 3.5 280777

Con la proximidad del mundial de básquet de 2019 y considerando que para replicar el paper de Song [SZS18] contado brevemente antes se necesitaban una gran cantidad de datos difíciles de conseguir, se pensó en adaptar el modelo usado en 301060 para predecir los resultados de este torneo internacional y así nació 280777.

#### 3.5.1 Modelado

Durante el proceso de adaptación del modelo nos encontramos con varios problemas que se pueden esperar dada la gran diferencia que existe entre estos deportes en términos de la cantidad de conversiones obtenidas por los equipos. Tomando una base de datos de más de 4500 partidos podemos empezar a analizar esta problemática.




En el histograma anterior podemos visualizar nuestra base de datos donde se muestra que las medias de los puntos en los partidos de básquet se encuentran entre los 70 y 90 puntos mientras que en el fútbol es menor a 3 goles.
Un problema fácil de ver, producto del modelo a utilizar, es que la distribución de poisson tiene media y varianza iguales, por lo tanto, esto les otorga probabilidades altas, mayores al 5%, a resultados que son muy extraños, cómo hacer menos de 40 puntos, algo que sólo encontramos 3 veces en la base de datos (2 iguales a 40), y casos de puntuaciones mayores a los 120 que sólo ocurre en 7 encuentros de los utilizados para entrenar el modelo. Además las medias y varianzas tan grandes generan diferencias irreales entre los equipos para el modelo, otorgando probabilidades mayores al 95% a favor del que considera más fuerte.
Por último los partidos de básquet no pueden terminar empatados, se juegan tiempos extras hasta que alguno logre la victoria.
Teniendo en cuenta estas diferencias que se notan al analizar los distintos deportes aplicamos las siguientes modificaciones:

- **Poner un piso de 40 puntos:** La solución pensada para el primer problema consta de suponer que los equipos anotarán un mínimo de 40 puntos. Para lograr que el modelo refleje esta suposición lo que se hizo fue restarles 40 puntos a todos los resultados de la base, en los 3 partidos en los que el equipo no llegó a ese puntaje crítico se consideró que sí los obtuvo. Al puntaje que el modelo prediga sobre un equipo se le sumará los 40 puntos descontados.

- **Disminuir y juntar varianzas mediante un factor de 5:** Para reducir las probabilidades de los puntajes por encima de los 120 y poder lograr que los porcentajes de probabilidades de victorias sean acordes a la realidad se buscó reducir las varianzas. Para conseguir esto se propuso dividir los parámetros calculados por un factor que luego será utilizado para multiplicar el resultado devuelto por la distribución de Poisson. Se determinó que 5 sea ese factor luego de que dicho número sea el que asemeje mejor la media de aciertos del modelo con los aciertos reales. En la tabla de abajo se muestran los valores obtenidos en los datos de entrenamiento.

- **Dar 50/50 de probabilidad de ganar el partido:** Como los casos en que los equipos terminan empatados son muy pocos como para poder predecir bien los tiempos extras, y si llegaron a ese resultado es que los dos equipos están teniendo un nivel muy similar en ese partido específico, lo que vamos a asumir es que la probabilidad de ganar el partido en tiempo extra sea del 50% para cada equipo.

Para realizar la siguiente tabla, tomamos los 2 años previos a cada competición el Mundial de China y Super 20 2019. Los porcentajes de acierto del modelo fueron 69% (779 sobre 1129 posibles) para los equipos de la Liga nacional y 67% (239 de 353) para las selecciones.

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| Esperanza Selecciones | 87% | 83% | 79% | 76% | 74% | 72% | 70% | 69% | 67% | 66% |
| Esperanza Liga | 73% | 67% | 65% | 63% | 62% | 60% | 59% | 59% | 58% | 57% |
| Error cuadrático medio | 377 | 229 | 137 | 100 | 85 | 97 | 104 | 101 | 122 | 148 |

Cómo el factor a definir sólo lo usamos para dividir al parámetro lambda de todos los equipos por igual, no afecta al favorito del partido, sólo a la probabilidad de ganar que se le otorga. Luego, denominando esperanza al porcentaje de los partidos predichos a los cuales esperamos dar con el resultado real, buscamos al factor que mejor ajuste la esperanza obtenida con el porcentaje de partidos en los que realmente encontró al ganador de acierto del modelo.
En conclusión, combinamos los errores utilizando la fórmula del error cuadrático medio, llegando a la determinación de que el número que estábamos buscando es 5.

Juntando los items previos, el puntaje final que nosotros le otorgamos a los equipos $X$ e $Y$ que se enfrenten en un partido, siendo $P_X$ y $P_Y$ el resultado indicado por sus respectivas distribuciones de Poisson serán de la forma $P_X * 5 + 40$ y $P_Y * 5 + 40$.

En lo que sigue se explican qué parámetros calculamos y la forma en la que lo hacemos, diferenciando lo hecho para el mundial con lo hecho para los equipos de la liga nacional.

## Mundial de China 2019

Después del procesamiento de los resultados de los partidos previos, recolección y reducción de 40 puntos a todos los puntajes (si no llega a 40, se asume que si llegó), debemos calcular los parámetros que necesitamos siguiendo los pasos del paper [DC97], la fuerza de ataque y defensa de todos los equipos y los agregados por la modificación para el sitio 301060 el plus de ataque y defensa local que en este caso sólo es necesario para China, el país anfitrión del torneo mundial. Para esto se usará una regresión lineal de Poisson a partir de la siguiente función de pseudo-probabilidad,

$$\prod_{k=1}^{N}[(\lambda_{ik}^{x_k} * e^{-\lambda_{ik}}) * (\mu_{jk}^{y_k} * e^{-\mu_{jk}})]^{e-0.0065(t-t_k)}$$

Donde $N$ es la cantidad de partidos de la base de datos, $X_k$ son los puntos marcados por el equipo local ($L_k$) y $y_k$ son los puntos marcados por el equipo visitante ($V_k$) en el partido $k$, $t_k$ la semana en la que se disputó dicho encuentro y $t$ la semana en la que se realiza la predicción.
Luego los parámetros calculados son;
$\lambda_k = exp(AtL_k - DefV_k + PAL_k)$ y $\mu_k = exp(AtV_k - DefL_k - PDL_k)$ tal que $AtL_k$ y $DefL_k$ son la fuerza de ataque y defensa del equipo local, $PAL_k$ y $PDL_k$ son el plus de ataque y defensa de la localía y $AtV_k$ y $DefV_k$ son la fuerza de ataque y defensa del equipo visitante.

Notar que aunque para las predicciones de este mundial sólo usaremos los parámetros de localía de China calcularemos todos para, de esta forma, no perder la diferencia que se nota en los resultados previos entre jugar de local, visitante o en cancha neutral. En los casos donde el partido se jugó en cancha neutral se toman el plus de ataque ($PAL_k$) y de defensa $PDL_k$ iguales a 0.

A cada partido se le asigna un peso según la fecha que se jugó utilizando como función decreciente $\phi = exp(-0.0065(t - tk))$ siendo $t$ el día cuando se entrena el modelo, $tk$ la jornada cuando se jugó el partido y $t - tk$ la diferencia en semanas entre las fechas anteriores. El parámetro de la función $\phi$ se seleccionó al ser el que mejor predecía en base a una muestra de entrenamiento.

Cómo en éste deporte algunas selecciones no disputan todas las competencias con el mejor equipo, en cambio sí lo suelen hacer en los mundiales y Juegos Olímpicos, se tomó la decisión de que a todos los partidos de estas dos competencias se consideren cómo si se hubieran jugado 3 años después, para así darle mayor importancia en la regresión de Poisson que a un partido amistoso.

Liga Nacional y Super 20

En este caso, también realizaremos el procesamiento de la base de datos a utilizar y se busca obtener la fuerza de ataque, fuerza de defensa, plus de ataque y plus de defensa de todos los equipos.
Para esto se usó una regresión lineal de Poisson a partir de la siguiente función de pseudo-probabilidad,

$$\prod_{k=1}^{N}[(\lambda_k^{x_k} * e^{-\lambda_k}) * (\mu_k^{y_k} * e^{-\mu_k})]^{e-0.008(t-t_k)}$$

Donde $N$ es la cantidad de partidos de la base de datos, $X_k$ son los puntos marcados por el equipo local ($L_k$) y $y_k$ son los puntos marcados por el equipo visitante ($V_k$) en el partido $k$, $t_k$ la semana en la que se disputó dicho encuentro y $t$ la semana en la que se realiza la predicción.
Luego los parámetros calculados son;
$\lambda_k = exp(AtL_k - DefV_k + PAL_k)$ y $\mu_k = exp(AtV_k - DefL_k - PDL_k)$ tal que $AtL_k$ y $DefL_k$ son la fuerza de ataque y defensa del equipo local, $PAL_k$ y $PDL_k$ son el plus de ataque y defensa de la localía y $AtV_k$ y $DefV_k$ son la fuerza de ataque y defensa del equipo visitante.

A cada partido se le asignó un peso según la fecha que se jugó utilizando como función decreciente $\phi = exp(-0.008(t - tk)$ siendo $t - tk$ siendo $t - tk$ la diferencia en semanas entre el día que se realiza la predicción y el día que se jugó el partido. En este caso también el parámetro de la función fue el ganador candidato para nuestra base de datos de entrenamiento apropiada, se puede notar que el exponente negativo resulta mayor que el de selecciones. Esto puede ser atribuido a que los equipos suelen cambiar en gran medida sus formaciones, y por lo tanto los partidos viejos interesan menos a la hora de predecir, al contrario que con las formaciones de los seleccionados se mantienen más estables.

### 3.5.2 Recolección y procesamiento de datos

Lo primero a buscar es una buena base de datos;
Para calcular los distintos parámetros de las distribuciones utilizamos resultados de torneos anteriores de los distintos equipos en los últimos años, la cantidad de años depende de las competiciones. La diferencia se nota entre los partidos internacionales y los nacionales debido a que el volumen de partidos anuales es mucho menos para las competencia de los países.

En los items siguientes se separará lo hecho con las selecciones para el Mundial de China y con los equipos de la Liga Nacional de Básquet.

### Mundial de China 2019

En el caso del mundial de China se tomaron como datos los resultados de los partidos de las últimas competiciones jugadas por las selecciones que competirían en el Mundial de China. Estos fueron el Mundial 2014, los Juegos Olímpicos 2016, las eliminatorias para el mundial que se estaba por jugar y las copas Continentales jugadas en 2015 y 2017 remarcando cuál de los equipos había jugado de local los partidos. En el caso de las competencias que se juegan completamente en un país los equipos que jugaban fueron denotados como neutrales en todos sus partidos con la excepción de los encuentros donde jugó el anfitrión, pues éste juega de local y en consecuencia, sus rivales serán visitantes. Es importante remarcar que se toman la totalidad de éstos partidos, aunque en la mayoría esté implicada alguna selección que no está clasificada al evento que se va a predecir, esto lo hacemos para no perder datos y poder ajustar mejor los parámetros a calcular. Por último se eliminaron de la base de datos algunos partidos atípicos porque se conocía que alguna de las selecciones implicadas jugó el encuentro con un equipo inferior al que luego disputaría el mundial.

### Liga Nacional y Super 20

En el caso de los torneos de la liga de básquet nacional se tomaron como datos los resultados de los partidos de las ediciones de primera división, Liga Nacional y Super20 de los últimos tres años. También se incluyeron a esa base los torneos de la Liga Argentina, segunda división del básquet nacional, de los mismos años para poder estimar el nivel de los equipos recién ascendidos. Hacer esto puede sobrevalorar a los equipos recién ascendidos, pues para lograrlo vienen de realizar una muy buena campaña, pero no sucede gracias a los equipos que estuvieron cambiando de categoría en esos años, ya que al jugar con los equipos de ambas divisiones las fuerzas de todos los participantes se logran nivelar. Igual que con los internacionales se diferenció al equipo local.

### 3.5.3 Predicción y simulación de los torneos

Para la predicción de los partidos se buscará estimar la probabilidad de que un equipo convierta más cantidad de puntos que los contrincantes. Para conseguir lo deseado, usaremos las distribuciones de poisson con los parámetros conseguidos como se dijo anteriormente.
Por ejemplo, para la probabilidad de que el equipo A le gane al equipo B (notada $P(GA)$ ), siendo $X \sim P(\lambda)$ e $Y \sim P(\mu)$ sus distribuciones, buscaremos $P(Y < X)$ y para simplificar, usaremos que los equipos hacen entre 40 y 140 puntos, recordando que en nuestro modelo los puntos de un

equipos son iguales a X*5 + 40, luego $X \leq 20$ e $Y \leq 20$. En síntesis;

$$P(GA) = P(Y < X) = P(X = x, Y = y \ : \ 0 < x \leq 20 \, , \ 0 \leq y < x)$$

Para conseguir las probabilidades de los torneos no podemos conformarnos con la probabilidad de que un equipo gane debido a que si varios participantes de un grupo en fase de grupos de una copa o en toda una liga, ganan la misma cantidad de partidos se encontraran empatados y se definirán sus posiciones según los puntos que convirtieron y recibieron en los encuentros terminados.
Cómo conseguir las probabilidades que puede tener cada resultado es muy complejo, tiene muchos casos, los partidos serán simulados. Con esto queremos decir que se generará un valor aleatorio con las distribuciones de los dos equipos que se enfrentan y se asumirá que los valores devueltos son los puntos que anotaron cada uno en dicho partido. Predecir quién será el campeón de un torneo es una tarea todavía más difícil.

Para las simulaciones de cada torneo se llevarán a cabo 1.000.000 de iteraciones. Este número se ha determinado debido a que, después de alcanzar esta cantidad de repeticiones, las diferencias en los resultados entre diversas simulaciones se sitúan en el orden del tercer decimal.

## 3.6 Programación

En esta sección se muestran los pseudocódigos de los algoritmos de predicción desarrollados y utilizados en esta tesis, los cuales fueron escritos en el lenguaje de programación R.

### 3.6.1 Probabilidad de partidos

Calcular los parámetros $\lambda$ y $\mu$ de las distribuciones del equipo local y visitante respectivamente
Armar matriz de las probabilidades de los distintos puntos de la distribución de Poisson conjunta $P(\lambda) * P(\mu)$ tal que $P(\lambda) \leq 20$ y $P(\mu) \leq 20$
Crear, usando la matriz de arriba, un resultado en forma de lista de 11 posiciones de la siguiente manera:
resultado[1]= Suma de todas las probabilidades donde el equipo local gana por 5 puntos
resultado[2]= Suma de todas las probabilidades donde el equipo local gana por 10 puntos
resultado[3]= Suma de todas las probabilidades donde el equipo local gana por 15 o y 20 puntos
resultado[4]= Suma de todas las probabilidades donde el equipo local gana por 25 y 30 puntos
resultado[5]= Suma de todas las probabilidades donde el equipo local gana por más de 30 puntos
resultado[6]= Suma de todas las probabilidades donde los equipos empatan en los 40 minutos
resultado[7]= Suma de todas las probabilidades donde el equipo visitante gana por 5 puntos
resultado[8]= Suma de todas las probabilidades donde el equipo visitante gana por 10 puntos
resultado[9]= Suma de todas las probabilidades donde el equipo visitante gana por 15 y 20 puntos
resultado[10]= Suma de todas las probabilidades donde el equipo visitante gana por entre 25 y 30 puntos
resultado[11]= Suma de todas las probabilidades donde el equipo visitante gana por más de 30 puntos
Sumar la probabilidad de empate dividida a la mitad a las probabilidades de ganar por 5 goles tanto para el local cómo para el visitante.
Devolver le vector resultado

### 3.6.2 Simulación: Mundial de China 2019

Para que nuestro algoritmo de simulación del torneo mundial tenga sentido, tiene que recrear exactamente el recorrido de todos los equipos participantes según el formato utilizado en esta competición.

La Copa consiste de una fase de grupos, donde los 32 equipos se dividen en 8 grupos de 4 equipos cada uno. Se enfrentan todos contra todos dentro de su grupo, avanzando los dos mejores a la segunda ronda, donde conforman grupo con otros 2 países a los que enfrentan 1 vez para luego avanzar los dos mejores de cada grupo a la siguiente fase.
En la fase de play-offs, empezando por cuartos de final, los equipos se enfrentan en partidos únicos, avanzando los ganadores y quedando eliminados los perdedores consiguiendo así determinar el campeón del mundo. Los equipos eliminados en cuartos de final disputan el quinto puesto en dos rondas eliminatorias.

Código
Crear los contadores de llegar a cada instancia, campeón, final, etc., para los 32 países
**Para** torneo=1 **hasta** 1000000:
    Simular todos los partidos de los grupos
    Analizar estos resultados para ordenar los grupos*
    Crear los nuevos grupos con los países que consiguieron el primer o segundo lugar y sumar 1 en los contadores de pasar a la segunda ronda de estos equipos
    Simular todos los partidos de los grupos
    Analizar estos resultados para ordenar los grupos*
    Crear las llaves de cuartos de final con los países que consiguieron el primer o segundo lugar y sumar 1 en los contadores de pasar a esta instancia.
    Simular los partidos de cuartos de final
    Ubicar en sus respectivas llaves de semifinal y sumar 1 en los contadores de pasar a esta instancia a los ganadores de cada partido
    Simular los partidos de semifinal
    Ubicar a los perdedores en el partido por el tercer puesto y a los ganadores en la final
    Sumar 1 en los contadores de pasar a la final a los ganadores
    Simular el partido por el tercer puesto
    Sumar 1 en el contador de tercer puesto del ganador
    Simular la final
    Sumar 1 en el contador de campeón del ganador
Fin **Para**
Dividir los contadores por la cantidad de repeticiones(1000000)
Guardar estos porcentajes en un archivo

### 3.6.3 Simulación: Super 20 2019

Esta copa consiste en una fase regular, donde los 20 equipos se dividen en cuatro grupos según ubicación geográfica, se enfrentan todos contra todos dentro de su grupo dos veces, una como local y otra como visitante. Los dos mejores equipos de cada grupo avanzan a los play offs.
La primera eliminatoria de los play offs es entre los equipos ubicados 1° y 2° de las zonas y es una serie al mejor de tres partidos, jugando el primero en cancha del peor ubicado y los dos restantes en la otra cancha.
Los ganadores acceden al Final Four, donde se ordenan según la marca obtenida en la fase de grupos, y se emparejan 1° contra 4° y 2° contra 3° donde los ganadores disputan la final. Esta etapa se disputa en una sede fija Y a partido único.

Código

Crear los contadores de llegar a cada instancia, campeón, final, etc., para los 20 equipos
**Para** torneo=1 **hasta** 1000000:
    Simular todos los partidos de los grupos
    Analizar estos resultados para ordenar los grupos*
    Sumar 1 en los contadores de pasar a cuartos de final de los equipos que consiguieron salir primeros o segundos
    Crear las llaves de cuartos de final con los equipos que consiguieron el primer o segundo lugar
    Simular los partidos de cuartos de final
    Ordenar los equipos que pasaron al Final Four según los partidos ganados en fase de grupos y sumar 1 en los contadores de estos equipos
    Crear las llaves de semifinal
    Simular los partidos de semifinal
    Ubicar a los perdedores en el partido por el tercer puesto y a los ganadores en la final
    Sumar 1 en los contadores de pasar a la final a los ganadores
    Simular el partido por el tercer puesto
    Sumar 1 en el contador de tercer puesto del ganador
    Simular la final
    Sumar 1 en el contador de campeón del ganador de la final
Fin **Para**
Dividir los contadores por la cantidad de repeticiones (1000000)
Guardar estos porcentajes en un archivo

### 3.6.4 Simulación: Liga Nacional

El formato de La Liga consta de dos fases, la fase regular y los play-offs:

**Fase regular:** los veinte equipos se enfrentan los unos a los otros dos veces, una vez como local y otra vez como visitante. Cada equipo que gane el partido obtiene dos puntos, mientras que aquel que pierda obtiene un punto. Con base a esos resultados se ordenan los equipos de mayor a menor puntaje obtenido en una única tabla.
Los doce mejores equipos, del 1° al 12°, avanzan al play-off para definir al campeón del torneo, mientras que el último y anteúltimo definen entre sí al peor equipo de la temporada y quien pierde la plaza para la siguiente temporada. Los restantes equipos dejan de participar, conservando su plaza para la próxima edición.

**Play-offs:** esta etapa se divide en 2 competencias separadas;
Permanencia los equipos ubicados en los puestos 19 y 20 definen en una serie al mejor de cinco partidos quien desciende. La serie se juega con formato 2-2-1, siendo local el equipo ubicado 19° en los primeros dos partidos y en el eventual último encuentro. Aquel equipo que pierda la serie queda relegado a la segunda división.

Campeonato: los cuatro mejores acceden a los cuartos de final, mientras que los puestos entre el 5 y el 12 disputan la reclasificación, al mejor de tres encuentros, enfrentándose 5° contra 12°, 6° contra 11°, 7° contra 10° y 8° contra 9°. Los cuatro vencedores de la reclasificación se reordenan en función de la posición en la tabla de la fase regular de manera que se enfrentan en cuartos de final según el siguiente ordenamiento: 1° contra peor clasificado, 2° contra segundo peor clasificado, 3° contra segundo mejor clasificado, 4° contra mejor clasificado. Esta instancia se disputa, cómo el resto de los play-offs, al mejor de cinco partidos y los ganadores de las series acceden a las semifinales. Los ganadores de las semifinales disputan la final, donde el equipo que venza en la serie se proclama campeón de la temporada.

Código

Crear los contadores de llegar a cada instancia, campeón, play-off, etc., para los 20 equipos
**Para** torneo=1 **hasta** 1000000:
    Simular todos los partidos de la fase regular (todos contra todos, ida y vuelta)
    Calcular las posiciones de la liga*
    Sumar 1 en los contadores de ingresar a la reclasificación de los equipos en los puestos entre el 5 y el 12
    Sumar 1 en los contadores de partido por el descenso de los últimos dos equipos
    Simular la serie por el descenso entre estos equipos***
    Sumar 1 en el contador de descenso del equipo derrotado en la serie
    Simular los partidos de reclasificación
    Sumar 1 en los contadores de ingresar a play-off de los equipos que consiguieron las primeras 4 posiciones y los ganadores de su partido de reclasificación
    Crear los cuartos de final con los equipos que pasaron a play-off
    Simular las series de cuartos de final***
    Ubicar en sus respectivas llaves de semifinal y sumar 1 en los contadores de pasar a esta instancia a los ganadores
    Simular las series de semifinal***
    Ubicar a los perdedores en el partido por el tercer puesto y a los ganadores en la final
    Sumar 1 en los contadores de pasar a la final a los ganadores
    Simular el partido por el tercer puesto
    Sumar 1 en el contador de tercer puesto del ganador
    Simular la final***
    Sumar 1 en el contador de campeón del ganador de la final
Fin **Para**
Dividir los contadores por la cantidad de repeticiones (1000000)
Guardar estos porcentajes en un archivo

*Se ordenan por partidos ganados, si se tienen varios empatados se define por cantidad de partidos ganados entre ellos y de seguir la paridad por diferencia de puntos en estos partidos.

**Las serie de reclasificación se juegan al mejor de 3 partidos. Se juega el primer partido con el equipo que termino peor en la tabla general de local y los 2 restantes cambiando de localía.

***Las series de cuartos de final, semifinal, final y descenso se juegan al mejor de 5 partidos. Se juegan 3 con el equipo que termino mejor en la tabla general de local y los restantes de visitante.

# 4 Resultados Obtenidos

En esta sección se mostrarán los resultados de las predicciones realizadas en los torneos, Mundial de básquet 2019 y el Super20 2019 y temporada 2019-2020 de la Liga Nacional.

Para testear los distintos modelos utilizaremos distintas estrategias que son bastante intuitivas de pensar al querer probar la efectividad de algoritmos de predicción.
Cómo primer testeo compararemos los porcentajes de pasar a las distintas etapas que dio inicialmente el modelo en los torneos a analizar, contra lo que ocurrió luego. Hay que tener en cuenta que algunos equipos por más de que tuvieran muchas chances de pasar, por ejemplo a cuartos de final, se encontraban en un grupo con otros 2 con mayores o las mismas posibilidades de pasar y por el diseño del torneo sólo pasarán 2.

Luego ya más detallado veremos cada partido en particular, contabilizando los encuentros en los cuales el modelo ”acertó” el resultado.
¿Qué se quiere decir con que el modelo acertó el resultado? Tomaremos como futuro ganador al equipo al que el modelo le otorgue el mayor porcentaje de ganar el partido a jugarse y luego contaremos los partidos en los que este ”futuro ganador” se llevó realmente el encuentro.
Por último vamos a tratar de ser un poco más justos con nuestro predictor, pues si un equipo tiene un 60% de ganar el partido, a su vez tiene un 40% de perderlo y aunque esto último es obvio, no se tiene en cuenta en las otras métricas. Lo ideal seria que se juegue muchas veces el partido para así ver exactamente el porcentaje de victorias de cada equipo, pero como esto es imposible, lo que haremos es juntar todos los partidos con probabilidades parecidas de que el favorito gane y comparar con el porcentaje de victorias de estos equipos en esos partidos.

## 4.1 Mundial de China 2019

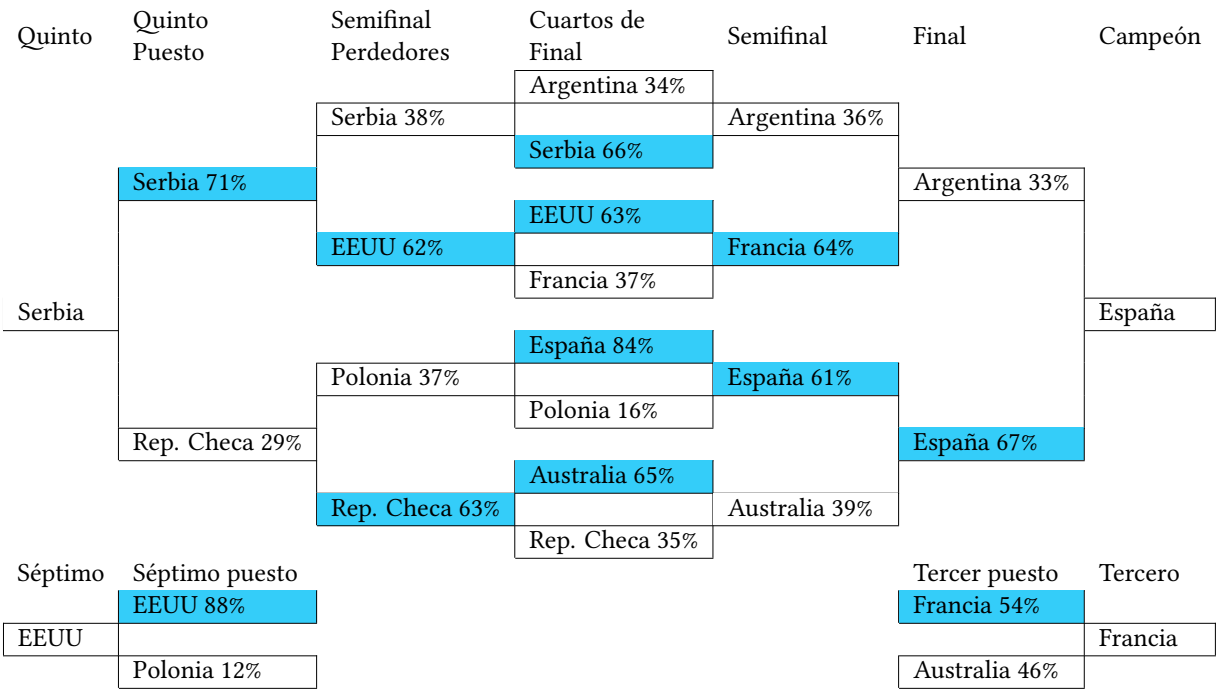
En la siguiente tabla pueden verse las probabilidades otorgadas por el modelo de llegar a cada una de las fases eliminatorias de los ocho principales candidatos y de los dos ”infiltrados” en los cuartos de final con su numero de orden en la lista completa de países.
Las casillas verdes indican que el equipo llegó a esa instancia y el número que acompaña a los nombres representa la posición en el ranking del modelo.

| Países | Cuartos de final | Semifinal | Final | Campeón |
|---|---|---|---|---|
| 1. Estados Unidos | 90.19 | 65.05 | 46.54 | 31.77 |
| 2. España | 84.68 | 68.84 | 42.45 | 25.23 |
| 3. Serbia | 72.84 | 52.53 | 25.53 | 11.26 |
| 4.Francia | 58.29 | 30.92 | 16.73 | 8.02 |
| 5. Australia | 44.76 | 22.27 | 11.52 | 5.50 |
| 6.Grecia | 41.15 | 19.75 | 9.19 | 3.51 |
| 7.Alemania | 40.00 | 17.61 | 8.28 | 3.34 |
| 8. Argentina | 60.18 | 23.73 | 5.90 | 2.86 |
| 14.República Checa | 20.75 | 7.39 | 2.62 | 0.69 |
| 15. Polonia | 37.78 | 8.98 | 1.98 | 0.40 |

De esta gráfica observemos que el campeón finalmente fue España, el segundo país favorito para nuestro modelo con un un poco más del 25%.
Analizando esta tabla podría decirse que le fue muy bien, pues sacando a Argentina, que se podría decir que fue la sorpresa del mundial, dejando a afuera a Serbia y Francia (tercer y cuarto país en la pelea por el título). Pues encontró a 7 de los 8 participantes de cuartos de final, República Checa se ubicó por encima de Grecia. El caso de Polonia es distinto, pues era el mas probable de su grupo mientras que Alemania se encontraba en el grupo con otros dos mejores puntuados, Australia y Francia.

Ahora miremos cómo le fue al modelo en cuestión de partidos acertados.
De los 48 partidos de la primera fase de grupos, en 42, el 87.5% de estos, ganó el equipo que más probabilidades tenía de ganar según 280777.
De los 32 partidos de la segunda fase de grupos y los partidos de los puestos 17 al 32, en 26 el 81.25%, ganó el equipo que más probabilidades tenía de ganar según el modelo.
En la etapa eliminatoria, de los 12 partidos jugados, en 8 el 66.67% pasó de ronda el favorito.
Por lo tanto, el modelo acertó en 76 partidos de los 92 jugados en todo el torneo, el 82,61%.

A continuación se puede ver la llave completa con los porcentajes asignados por encuentro y resaltados nuestros favoritos.




Por último veamos como que tan bien aproximadas están las probabilidades que devuelve nuestro modelo.

| Porcetaje 280777 | Cantidad de Partidos | Cantidad Acertados | Porcentaje Promedio | Aciertos |
|---|---|---|---|---|
| Mas de 75% | 34 | 34 | 85.26% | 100% |
| Entre 65% y 75% | 21 | 18 | 70% | 85.72% |
| Entre 60% y 65% | 17 | 12 | 63.12% | 70.59% |
| Entre 60% y 55% | 9 | 5 | 57.11% | 55.56% |
| Entre 55% y 50% | 11 | 7 | 53% | 63.64% |
| TOTAL | 92 | 76 | 71.08% | 82.61% |

En la tabla podemos ver que la mayoría de los partidos tenían una probabilidad por arriba del 65% para el equipo favorito y que el modelo en estos suele subestimar las probabilidades, es decir, los equipos con porcentajes tan altos suelen ganar sus partidos, se ve mejor en los que obtuvieron más del 75% donde los 34 partidos fueron predichos correctamente.
Luego con los partidos mas parejos, los porcentajes se encuentran mas cercanos a lo ocurrido en el torneo, es más, los partidos con porcentajes entre 60 y 55 fueron sobrestimados por nuestro modelo. Viendo el total de los partidos, sin discriminar por porcentajes se nota una tasa de aciertos notablemente superior a la esperada por el modelo, debido a la alta cantidad de partidos con un favorito muy superior.

Para comparar un poco el modelo, utilizaramos el ranking FIBA (Federación Internacional de Baloncesto). El FIBA World Ranking es un sistema de clasificación para equipos nacionales de básquet, éste se calcula asignando puntos a los equipos según su desempeño en competiciones internacionales. La cantidad de puntos depende del resultado del partido, la importancia del torneo y la región del equipo. Los puntos se ponderan en función del tiempo, dándole más peso a los resultados recientes. También se aplican coeficientes continentales y se recompensa la participación continua en competiciones internacionales. Este sistema se actualiza regularmente para reflejar el rendimiento más reciente de los equipos.

Si para este torneo siempre apostamos al país mejor posicionado según el Ranking, se aciertan un total de 40 partidos, el 43,48%, lo que nos muestra que el método de realización del ranking no condice tanto con la realidad de los equipos.
Comprando esto con nuestras predicciones, en un total de 92 partidos en 43 acertó el modelo de 28077 en comparación con el ranking, en 7 fue al revés, en 33 los dos modelos predijeron correctamente al ganador del partido y en 9 la predicción fue errónea en ambos lados.
Viendo estos resultados, podríamos indicar que probablemente el ranking formulado por la FIBA no tiene una gran veracidad en cuanto al poderío de un equipo, una causa de esto podrían ser las pocas competencias internacionales que se realizan principalmente entre los distintos continentes, algo que es muy complejo de compensar sin un algoritmo.

### 4.2 Super 20

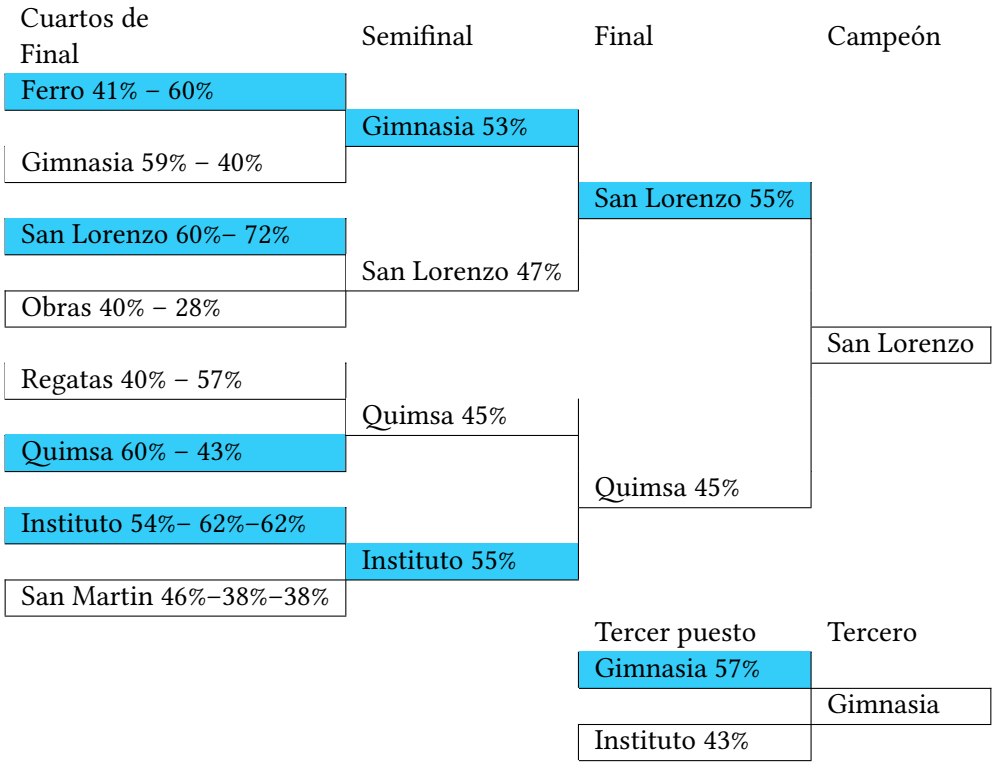
En la siguiente tabla pueden verse las probabilidades de llegar a las fases eliminatorias de los ocho principales candidatos y los cuales lograron ingresar a los cuartos de final del torneo.
Las casillas verdes indican que el equipo llegó a esa instancia y el número que acompaña a los nombres representa la posición en el ranking del modelo

| Equipos | Cuartos de final | Semifinal | Final | Campeón |
|---|---|---|---|---|
| 1. San Lorenzo | 81.71 | 44.94 | 29.73 | 19.46 |
| 2. Instituto | 61.28 | 31.37 | 18.14 | 10.18 |
| 3. Ferro | 58.83 | 29.97 | 15.60 | 7.93 |
| 4. San Martín | 50.05 | 25.29 | 13.16 | 6.68 |
| 5.Obras Sanitarias | 52.06 | 26.12 | 13.06 | 6.37 |
| 6.Gimnasia (CR) | 44.57 | 21.19 | 10.58 | 5.15 |
| 7. Quimsa | 42.94 | 21.43 | 10.70 | 5.14 |
| 8. Regatas | 40.03 | 19.95 | 9.96 | 4.81 |

De esta gráfica observemos que San Lorenzo terminó consiguiendo el título siendo éste el club al cual el modelo lo tuvo como favorito desde el comienzo con un 19.46% que no es mucho por lo parejo del torneo pero a su vez casi duplica las chances del que aparece como segundo.
También vale destacar que el modelo acertó los 8 equipos que llegaron a cuartos de final.

Ahora miremos cómo le fue al modelo en cuestión de partidos acertados.
De los 80 partidos de la fase de grupos uno no se jugó por problemas técnicos en los tableros de la cancha. De los 79 restantes el modelo acertó en 51 un 64.55%. La etapa de eliminación consiste de una fase al mejor de 3 partidos donde el primero se jugará en la cancha del equipo peor clasificado (segundo de su grupo) y los 2 siguientes con el otro equipo de local. En este caso, sólo Instituto contra San Martín llegó al tercer partido y de los 9 partidos de los que se jugaron en 6, el 66.67% ganó el equipo que más probabilidades tenía de ganar según 280777.
Las dos fases siguientes, semifinal y final, se juega a un sólo partido y todos en la cancha de Gimnasia, por lo tanto en los partidos que éste conjunto no juegue, ambos equipos serán neutrales, es decir, no contarán con plus de localía.
En las semifinales, nuestro modelo tuvo un pequeño tropiezo equivocándose en ambos, pero volvió a acertar el partido final y el tercer puesto cerrando un 50% en esta etapa.

A continuación se puede ver la llave completa con los porcentajes asignados por encuentro, en la etapa de cuartos 2 o 3 partidos, y resaltados nuestros favoritos en cada uno.



Por último veamos que tan bien aproximadas están las probabilidades que devuelve nuestro modelo.

| Porcetaje 280777 | Cantidad de Partidos | Cantidad Acertados | Porcentaje Promedio | Aciertos |
|---|---|---|---|---|
| Mas de 75% | 3 | 2 | 78.33% | 66.67% |
| Entre 75% y 65% | 14 | 13 | 69.43% | 92.86% |
| Entre 65% y 60% | 25 | 18 | 63.16% | 72% |
| Entre 55% y 60% | 23 | 16 | 57.22% | 69.57% |
| Entre 55% y 50% | 27 | 13 | 52.26% | 48.15% |
| TOTAL | 92 | 62 | 59.38% | 67.39% |

En el torneo local, los partidos suelen ser más parejos, por eso vemos que sólo 17 tienen un porcentaje por encima del 65% para el equipo favorito, pero el modelo sigue subestimando un poco las probabilidades de estos partidos,pues de los 17 partidos que cumplen esto en 15, un 88.24%, la predicción era correcta.

Analizando los otros grupos de partidos, nuevamente se ve una subestimación pero mucho más cercana a lo ocurrido.
Por último, viendo el total de los partidos, sin discriminar por porcentajes, la tasa de aciertos no quedaron tan lejanos al porcentaje de encuentros que esperaba predecir correctamente el modelo.

## 4.3 Liga Nacional

Veamos cómo le fue al modelo en cuestión de predicción del torneo en general. Vale notar que esta liga se debió suspender en la mitad por la cuarentena adoptada por Argentina en marzo del 2020 en razón de los casos de COVID 19, por lo que sólo se jugaron 248 partidos de la fase regular y cómo el fixture utilizado no es simétrico en cuestión de partidos, hubo equipos con hasta 26 partidos jugados mientras otros sólo 20 de los 38 que se juegan en un torneo completo.

En la siguiente tabla se pueden ver los primeros puestos del torneo jugado con las probabilidades que nuestro modelo les otorgó a cada uno de llegar a las distintas instancias o consagrarse como campeón del mismo

| Equipo | PG | PJ | PG% | CAMPEON | PRIMERO | PLAYOFF |
|---|---|---|---|---|---|---|
| 1. Quimsa | 17 | 20 | 85.7 | 6.71 | 5.66 | 64.43 |
| 2. San Lorenzo | 17 | 21 | 81.0 | 39.52 | 47.97 | 96.51 |
| 3. Gimnasia (CR) | 19 | 25 | 76.0 | 6.32 | 4.96 | 62.57 |
| 4. Instituto | 15 | 22 | 68.2 | 19.7 | 18.72 | 86.97 |
| 5. Comunicaciones | 16 | 26 | 61.5 | 2.76 | 2.16 | 45.21 |
| 6. Ferro | 14 | 24 | 58.3 | 6.06 | 5.07 | 62.09 |
| 7. San Martin | 13 | 23 | 56.5 | 4.37 | 3.44 | 54.69 |
| 8. Regatas | 14 | 25 | 52.0 | 3.67 | 3.11 | 52.04 |
| 9. Boca Juniors | 13 | 26 | 50.0 | 1.65 | 1.30 | 35.98 |
| 10. Olímpico | 12 | 24 | 50.0 | 1.45 | 1.16 | 34.29 |

Aún ocurriendo esto y comparando con la simulación del torneo completo puede notarse un claro dominio de Quimsa, San Lorenzo y Gimnasia con más del 75% de partidos ganados los cuales estaban tercero, primero y quinto respectivamente en la tabla de favoritos para 280777 de quedarse en la primera posición. Un poco más abajo, con el 68% de las victorias, se encuentra Instituto al cuál el modelo lo consideraba el segundo "mejor equipo" de la liga, es decir con más posibilidades del primer puesto, lo que sigue indicando que la simulación no estaba tan errónea en su predicción.
Otro dato a destacar es que exceptuando a Comunicaciones, el impredecible para el algoritmo usado, los primeros 8 obtuvieron porcentajes por arriba del 50% de ingresar a playoff, es decir de mantenerse entre los ocho mejores al terminar la fase regular.

Ahora comparemos los últimos puestos con los porcentajes otorgados al descenso y último puesto

| Equipo | PG | PJ | PG% | ULTIMO | DESCENSO |
|---|---|---|---|---|---|
| 11. Weber Bahía | 11 | 26 | 46.2 | 17.06 | 17.72 |
| 12. Hispano | 12 | 27 | 44.4 | 10.81 | 10.76 |
| 13. Argentino | 11 | 26 | 42.3 | 15.74 | 16.20 |
| 14. Obras Sanitarias | 11 | 26 | 42.3 | 1.41 | 1.27 |
| 15. Peñarol | 10 | 26 | 40.7 | 5.94 | 5.84 |
| 16. Atenas | 10 | 25 | 40.0 | 2.80 | 2.59 |
| 17. La Union | 10 | 25 | 40.0 | 1.73 | 1.55 |
| 18. Platense | 8 | 24 | 33.3 | 11.03 | 11.25 |
| 19. Estudiantes | 7 | 26 | 25.9 | 5.90 | 5.82 |
| 20 Libertad | 5 | 25 | 23.1 | 18.22 | 18.51 |

En este apartado puede verse que al final de la tabla se encuentra Libertad, equipo al que el modelo daba como favorito a quedar en esta posición pero luego notamos que los otros clubes con probabilidades altas se encuentran alejados de esa última posición.
Estos ”errores” se pueden originar debido a que los equipos más débiles no se llevan mucha diferencia de poderío entre ellos y por lo tanto, al modelo se le hace más difícil de predecir.

Miremos cómo le fue al modelo en cuestión de partidos acertados.
El torneo de la temporada 2019-2020 fue suspendido en la mitad por la cuarentena adoptada por Argentina en marzo del 2020 por los casos de COVID 19, por lo que solo se jugaron 248 partidos de la fase regular, entre los cuales acertó en 173 un 69.76%.

| Porcetaje 280777 | Cantidad de Partidos | Cantidad Acertados | Porcentaje Promedio | Aciertos |
|---|---|---|---|---|
| Mas de 75% | 9 | 9 | 79.67% | 100% |
| Entre 75% y 65% | 62 | 56 | 67.29% | 90.32% |
| Entre 65% y 60% | 40 | 29 | 60.2% | 72.5% |
| Entre 60% y 55% | 52 | 30 | 54.87% | 57.69% |
| Entre 55% y 50% | 85 | 49 | 50.78% | 57.65% |
| TOTAL | 248 | 173 | 58.33% | 69.76% |

Analizando la Liga, obtenemos mejores conclusiones gracias a la gran cantidad de partidos y la tendencia de subestimar las probabilidades de los encuentros con diferencias altamente marcadas, superiores al 60%.
Con los otros grupos de partidos, ya se ve una estimación más cercana, casi igualando los porcentajes en los encuentros con una probabilidad entre 60 y 55.
Por último, viendo todos juntos, sin discriminar por probabilidades, los partidos acertados no quedaron tan lejanos al porcentaje de aciertos que brindaba el modelo, la diferencia se debe a la gran cantidad de partidos con mucha diferencia.

## 4.4 Análisis global

| Porcetaje 280777 | Cantidad de Partidos | Cantidad Acertados | Porcentaje Promedio | Aciertos |
|---|---|---|---|---|
| Mas de 75% | 46 | 45 | 83.71% | 97.83% |
| Entre 75% y 65% | 97 | 87 | 68.19% | 89.69% |
| Entre 65% y 60% | 82 | 59 | 61.71% | 71.95% |
| Entre 60% y 55% | 84 | 51 | 55.75% | 60.71% |
| Entre 55% y 50% | 123 | 69 | 51.30% | 56.10% |
| TOTAL | 432 | 311 | 61.39% | 71.99% |

Por último haciendo un análisis global, es decir, salvando las pequeñas diferencias que tienen los modelos para cada torneo, podemos concluir que cumple su trabajo con alta efectividad, en un 71.99% de los partidos predichos ganó el equipo favorito.

También podemos atribuirle que puede diferenciar entre partidos parejos y partidos con un amplio favorito, en el 84.89% de los 225 partidos con porcentajes mayores al 60, el equipo al que se le atribuyó la victoria efectivamente ganó, mientras que en los de probabilidades menores, sólo fueron el 57.91% de los 207 encuentros que predijo.
Cómo una falencia, se le puede criticar que subestima las probabilidades de victoria de los equipos con grandes diferencias a sus rivales.

## 4.5 Comparación Song vs 280777

En esta sección se comparará el modelo de las dos Poisson independientes desarrollado por nosotros con el de la normal bivariada proveniente del paper [SZS18] ya explicado anteriormente.

Cómo ya comentamos es muy difícil juzgar un modelo probabilístico por si acertó o no un partido donde le otorgó 70% de probabilidad al ganador o si el equipo con más probabilidad de salir campeón logra ese objetivo pero al tener gran cantidad de partidos tendremos un mejor criterio para encontrar las ventajas y desventajas de cada uno.

Para comparar las predicciones de ambos algoritmos en de la temporada 2020-21 de la liga nacional, nos topamos con un problema, para poder utilizar el modelo desarrollado por Song se necesita entrenar con muchos datos, las estadísticas completas de los partidos (rebotes defensivos y ofensivos, pérdidas, cantidad de tiros de 2 y 3 puntos) datos que pudimos conseguir para las 2 temporadas anteriores de la primera división pero que no encontramos para torneos pasados de la segunda división del básquet local. Cómo se comenta en la sección **" Recolección y procesamiento de datos "** en la liga nacional todos los años asciende y desciende un equipo en la primera división, por lo que no poder tener las estadísticas necesarias de los encuentros de esta división genera errores para el entrenamiento de los modelos. En consecuencia, no podemos hacer un analisis completo de éste torneo pero si comparar los partidos jugados por 2 equipos que hayan permanecido en la primera, es decir, compararemos todos los partidos exceptuando en los que participe el último club ascendido, Oberá Tenis Club.

Para tener un análisis más completo para la comparación de estos modelos también utilizaremos la temporada 2021-22 del torneo de la NBA, competencia en la que se contabilizan en gran medida todas las estadísticas y que no cuentan con recambio de equipos, siempre juegan los mismo 30.

### 4.5.1 Comparación temporada 21-22 del torneo NBA

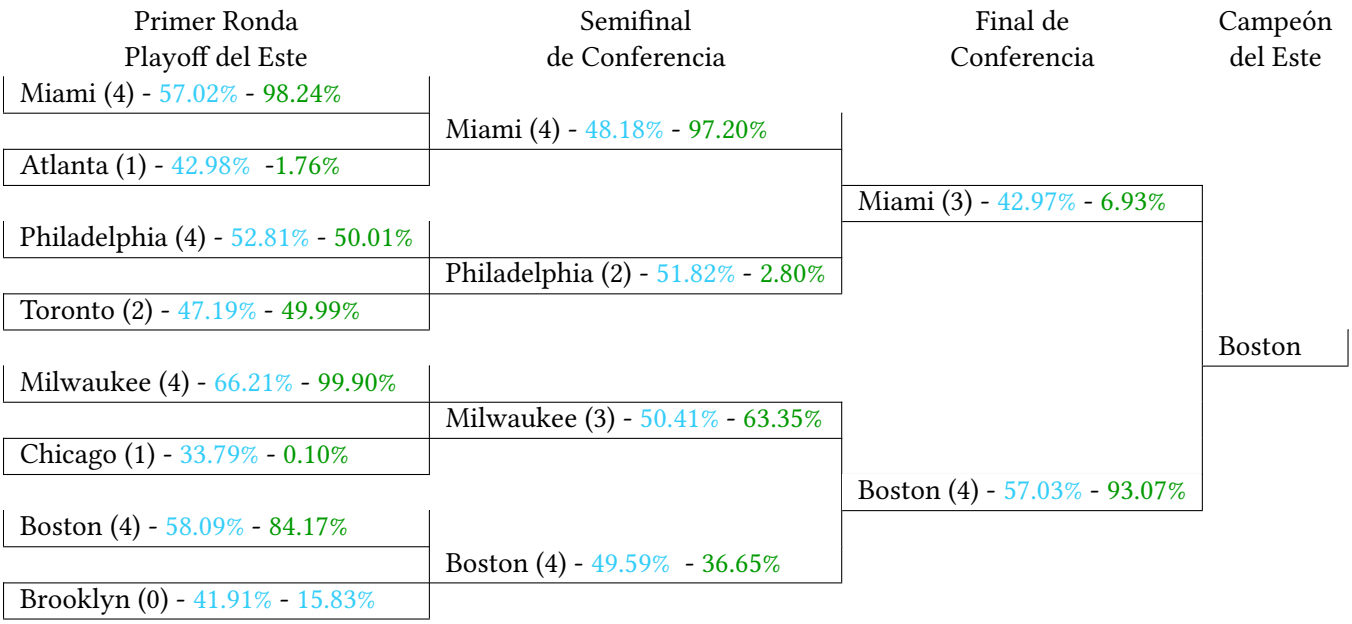
NBA es la abreviatura de ”National Basketball Association”, es la organización responsable del básquet profesional en los Estados Unidos, la competencia más importante de basquét del mundo. Esta liga está dividida en dos conferencias, Oeste y Este, cada una de las cuales tiene tres divisiones de cinco equipos. Durante una temporada regular, cada equipo juega 82 juegos, resultando en un total de 1230 partidos. Más específicamente, dos equipos en particular juegan 2 juegos si están en diferentes conferencias, un partido de local y el otro de visitante, jugando 4 juegos si están en la misma división, 2 en cada estadio, y 3 o 4 juegos si son en las diferentes divisiones de una misma conferencia, los casos de 3 partidos se van rotando mediante temporadas, siempre manteniendo que los 30 equipos jueguen 41 partidos de local y la misma cantidad de visitante. Para simpleza de la simulación supondremos que juegan 4 partidos con todos los de la misma conferencia, lo que agrega tan sólo 4 partidos extras en un total de 86.

Los mejores 6 equipos de cada conferencia clasifican a directo al play-Off y los otros 4 puestos de ésta eliminatoria, 2 de cada división, salen de una eliminación llamada play-In.
El Play-In consiste en dos series diferentes a un partido directo de eliminación. Primero, entre el 7º y 8º y después, entre el 9º y el 10º clasificado de cada Conferencia. El ganador del encuentro entre el 7º y 8º obtendrá la clasificación automática para los Playoffs de su respectiva Conferencia, en tanto, el perdedor de ese partido será local ante el vencedor del 9° y 10°, y el ganador de este nuevo encuentro será el octavo clasificado.

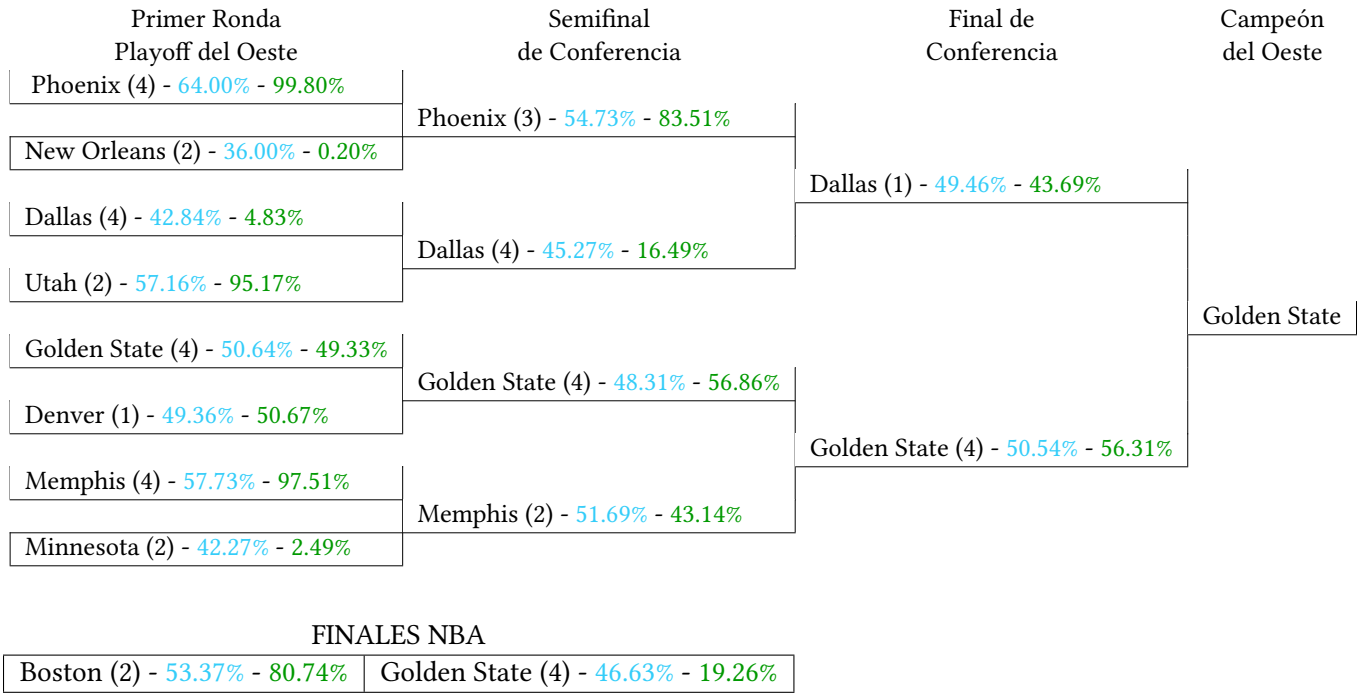
Una vez definidos los ocho clasificados de cada Conferencia, se juegan cuatro rondas al mejor de siete partidos, hasta coronar al campeón.
En la primera ronda, el primero de cada Conferencia se enfrenta con el octavo; el segundo, con el séptimo y así sucesivamente. Los ganadores avanzan a las Semifinales de Conferencia, luego a las Finales de Conferencia y eventualmente, a las Finales NBA.
En cada serie, la ventaja de campo es determinada por el récord de victorias y derrotas en la Fase Regular de cada equipo.

Como datos de entrenamiento utilizaremos todos los partidos que se jugaron desde el 1 de enero de 2019 hasta el final de la temporada 2020-2021 el 20 de julio de 2021, más de 3000 partidos.

| | Playoff % | | Cuartos % | | Semifinal % | | Final % | | Campeón % | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Equipo Este** | 280777 | Song | 280777 | Song | 280777 | Song | 280777 | Song | 280777 | Song |
| 1. Miami Heat | 93.89 | 98.90 | 22.20 | 4.81 | 4.70 | 0.19 | 0.77 | 0.01 | 0.12 | >0.01 |
| 2. Boston Celtics | 99.88 | 100 | 73.03 | 93.38 | 32.92 | 32.66 | 11.58 | 2.40 | 5.19 | 0.87 |
| 3. Milwaukee Bucks | 100 | 100 | 84.48 | 99.94 | 66.60 | 97.15 | 51.18 | 89.62 | 38.44 | 67.83 |
| 4. Philadelphia 76ers | 99.49 | 100 | 60.29 | 88.70 | 25.20 | 9.62 | 8.98 | 1.05 | 3.65 | 0.47 |
| 5. Toronto Raptors | 99.98 | 100 | 77.72 | 97.36 | 43.79 | 59.72 | 17.66 | 6.92 | 9.18 | 3.11 |
| 6. Chicago Bulls | 4.88 | 15.25 | 0.16 | 0.02 | 0.01 | >0.01 | >0.01 | >0.01 | >0.01 | >0.01 |
| 7. Brooklyn Nets | 96.16 | 99.72 | 30.89 | 11.61 | 8.10 | 0.48 | 1.86 | 0.01 | 0.43 | >0.01 |
| 8. Atlanta Hawks | 13.83 | 48.47 | 1.01 | 0.01 | 0.10 | >0.01 | >0.01 | >0.01 | >0.01 | >0.01 |
| 9. Cleveland Cavaliers | 0.1 | 0.02 | 0.01 | >0.01 | >0.01 | >0.01 | >0.01 | >0.01 | >0.01 | >0.01 |
| 10. Charlotte Hornets | 0.06 | 0.01 | >0.01 | >0.01 | >0.01 | >0.01 | >0.01 | >0.01 | >0.01 | >0.01 |



| | Playoff % | | Cuartos % | | Semifinal % | | Final % | | Campeón % | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Equipo Oeste** | 280777 | Song | 280777 | Song | 280777 | Song | 280777 | Song | 280777 | Song |
| 1. Phoenix Suns | 93.30 | 98.89 | 39.66 | 30.76 | 11.98 | 2.05 | 3.23 | 0.22 | 0.61 | 0.01 |
| 2. Memphis Grizzlies | 37.8 | 32.48 | 5.05 | 0.22 | 0.8 | 0.01 | 0.01 | >0.01 | >0.01 | >0.01 |
| 3. Golden State Warriors | 59.91 | 66.84 | 11.06 | 1.57 | 2.24 | 0.11 | 0.34 | 0.01 | 0.04 | >0.01 |
| 4. Dallas Mavericks | 79.18 | 86.02 | 20.94 | 10.92 | 4.91 | 0.96 | 1.01 | 0.04 | 0.14 | >0.01 |
| 5. Utah Jazz | 100 | 100 | 86.1 | 99.13 | 64.52 | 89.71 | 41.98 | 60.55 | 19.69 | 20.29 |
| 6. Denver Nuggets | 99.34 | 100 | 66.10 | 81.45 | 28.44 | 19.56 | 11.14 | 3.17 | 2.96 | 0.24 |
| 7. Minnesota Timberwolves | 0.1 | 0.01 | >0.01 | >0.01 | >0.01 | >0.01 | >0.01 | >0.01 | >0.01 | >0.01 |
| 8. New Orleans Pelicans | 6.34 | 1.74 | 0.50 | 0.02 | 0.04 | >0.01 | >0.01 | >0.01 | >0.01 | >0.01 |
| 9. Los Angeles Clippers | 99.96 | 100 | 86.08 | 96.57 | 63.27 | 81.05 | 36.22 | 35.17 | 14.68 | 7.16 |
| 10. San Antonio Spurs | 13.68 | 6.17 | 1.28 | 0.02 | 0.14 | >0.01 | 0.01 | >0.01 | >0.01 | >0.01 |




En las tablas anteriores se pueden ver las probabilidades de llegar a las distintas instancias de la postemporada que cada modelo le otorgó a los 10 mejores equipos de ambas conferencias y en las llaves se aprecia el resultado ocurrido en la serie al mejor de 7 partidos y los porcentajes que cada equipo tenía de ganarla, previo al encuentro propiamente dicho, distinguiendo con colores los valores del modelo desarrollado por nosotros en celeste y el extraído del paper de Song en verde

Como primer análisis de las tablas anteriores, se puede notar que ambos modelos consideran como poderosos a los mismos equipos, con la diferencia de que el de Song tiene valores mucho más extremos. Dicho esto, se puede considerar que en términos de acertar al campeón, Golden State Warriors, ambos estuvieron muy lejos (0.04% y >0.01%), mientras que los 5 favoritos eran Milwaukee Bucks (38.44% y 67.83%) perdió en semifinal de conferencia, Utah Jazz (19.69% y 20.29%),

Los Angeles Clippers (14.68% y 7.16%) no entró a los Playoff, Toronto Raptors (9.18% y 3.11%) perdió en primera ronda y Boston Celtics (5.19% y 0.87%) perdió en las finales.

Ahora hagamos un análisis global de estas predicciones;

### 280777

| Porcentaje 280777 | Cantidad de partidos | Acertados | Esperanza | Aciertos |
|---|---|---|---|---|
| Mas de 75% | 0 | 0 | - | - |
| Entre 65% y 75% | 10 | 5 | 66.44% | 50% |
| Entre 60 y 65% | 124 | 89 | 61.84% | 71.77% |
| Entre 55% y 60% | 410 | 270 | 57.07% | 65.85% |
| Entre 50% y 55% | 779 | 448 | 52.55% | 57.51% |
| TOTAL | 1323 | 812 | 54.93% | 61.38% |

## Normal bivariada (Song)

| Porcentaje Song | Cantidad de partidos | Acertados | Esperanza | Aciertos |
|---|---|---|---|---|
| Mas de 75% | 843 | 514 | 89.95% | 60.97% |
| Entre 65% y 75% | 208 | 124 | 70.31% | 59.62% |
| Entre 60 y 65% | 101 | 57 | 62.63% | 56.44% |
| Entre 55% y 60% | 81 | 46 | 57.81% | 56.79% |
| Entre 50% y 55% | 90 | 48 | 52.41% | 53.33% |
| TOTAL | 1323 | 789 | 80.26% | 59.64% |

En estas tablas agrupamos los partidos según el porcentaje de victoria que le otorgó el modelo al vencedor de los partidos.
A primera vista notamos que el modelo utilizado en 280777 es más precavido que el de Song, a la gran mayoría de los partidos se les dio un porcentaje menor al 60%, mientras que el otro modelo agrupa más de la mitad de los partidos con porcentajes que superan el 75%. esto se puede ver fácil en las esperanzas, donde el valor del primero es de 54.93% mientras que el segundo espera un 80.26%. Podemos elogiar el buen rendimiento que mantiene el modelo creado por nosotros con los partidos a los que les otorga un porcentaje mayor al 55%, acertando 364 de 544 partidos, un 66.91%.

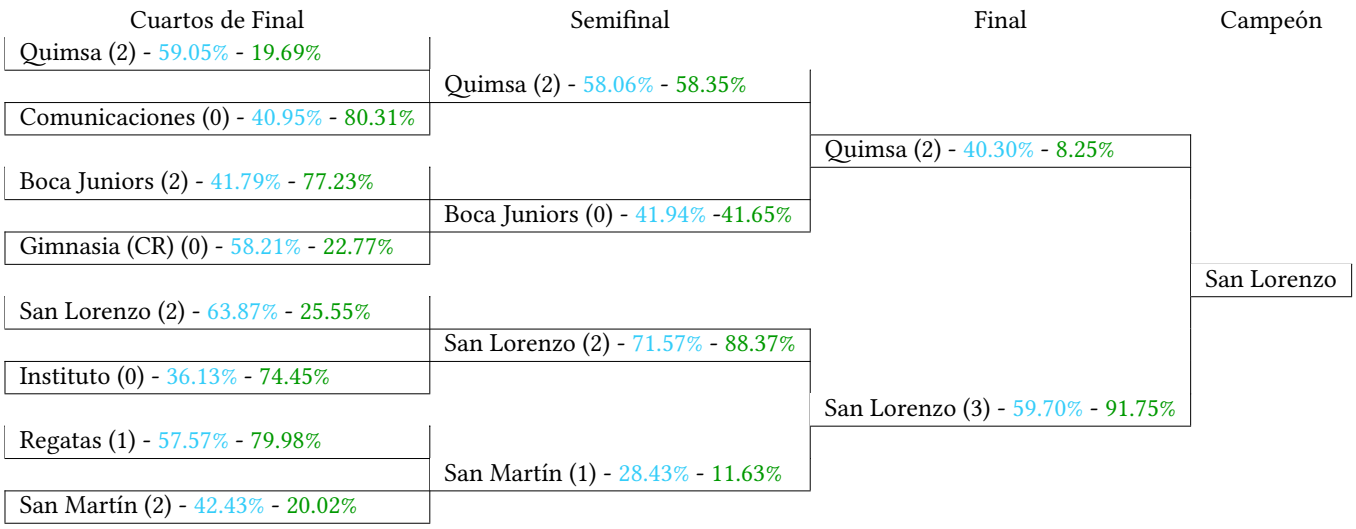
Comparando los modelos partido a partido se sigue notando la paridad en las predicciones. En un total de 1323 partidos en 77 acertó el modelo de 28077 y no el de Song, en 54 fue al revés, en 735 los dos modelos predijeron correctamente al ganador del partido y en 457 la predicción fue errónea en ambos algoritmos.

### 4.5.2 Comparación Torneo Nacional

En la temporada 2020-21 de la liga nacional se jugaron 411 partidos, en los cuales Oberá Tenis Club sólo participó en los 38 de la etapa regular, por lo cual usaremos como elementos de testeo los restantes 373 encuentros.
Por no poder predecir los partidos de todos los equipos, no es posible simular el torneo, pero considerando que en la temporada terminó en el puesto 17, simulamos el torneo como si se jugara con un equipo menos, para conseguir de todos modos los equipos mas fuertes para cada modelo.

| | Playoff % | | Cuartos % | | Semifinal % | | Final % | | Campeón % | |
|---|---|---|---|---|---|---|---|---|---|---|
| Equipo | 280777 | Song | 280777 | Song | 280777 | Song | 280777 | Song | 280777 | Song |
| 1. Quimsa | 93.30 | 99.79 | 39.66 | 50.29 | 11.98 | 11.40 | 3.23 | 1.70 | 0.61 | 0.09 |
| 2. San Lorenzo | 99.92 | 100 | 95.79 | 99.93 | 66.85 | 79.95 | 41.94 | 50.36 | 25.53 | 31.61 |
| 3. Regatas | 99.57 | 100 | 88.91 | 98.72 | 50.26 | 57.94 | 24.54 | 24.13 | 10.72 | 7.37 |
| 4. Boca Juniors | 94.97 | 99.87 | 60.44 | 61.90 | 22.36 | 13.78 | 7.48 | 2.45 | 2.02 | 0.18 |
| 5. Gimnasia (CR) | 97.31 | 99.98 | 71.56 | 83.01 | 31.63 | 29.59 | 12.75 | 8.95 | 4.31 | 1.18 |
| 6. Obras Sanitarias | 98.53 | 99.99 | 77.91 | 90.77 | 34.69 | 30.99 | 13.68 | 9.31 | 4.65 | 1.43 |
| 7. San Martín | 77.33 | 94.00 | 29.40 | 19.23 | 5.87 | 0.93 | 1.15 | 0.05 | 0.18 | >0.01 |
| 8. Instituto | 99.89 | 100 | 95.52 | 99.90 | 65.37 | 77.87 | 39.32 | 46.26 | 22.47 | 24.83 |
| 9. Olímpico | 50.10 | 58.71 | 13.02 | 1.60 | 2.24 | 0.04 | 0.38 | >0.01 | 0.05 | >0.01 |
| 10. Platense | 14.40 | 3.31 | 2.30 | 0.02 | 0.24 | >0.01 | 0.02 | >0.01 | >0.01 | >0.01 |
| 11. Comunicaciones | 99.87 | 100 | 94.47 | 99.91 | 63.63 | 74.74 | 39.32 | 51.41 | 22.47 | 32.47 |
| 12. Hispano | 51.07 | 52.36 | 12.91 | 1.85 | 2.00 | 0.05 | 0.30 | 0.01 | 0.04 | >0.01 |

| Primera fase de Playoff | |
|---|---|
| Obras Sanitarias (0) - 45.83% - 2.50% | Comunicaciones (2) - 54.17% - 97.50% |
| Gimnasia (CR) (2) - 70.74% - 99.40% | Hispano (0) - 29.26% - 0.6% |
| Instituto (2) - 63.43% - 99.01% | Olímpico (1) - 36.57% - 0.99% |
| San Mantín (2) - 56.56% - 99.10% | Platense (1) - 43.44% - 0.90% |




En la tabla anteriores se pueden ver las probabilidades de llegar a las distintas instancias de la postemporada que cada modelo le otorgó a los 12 mejores equipos del torneo, los ingresantes al Playoff, y en la llave se aprecia el resultado ocurrido en la serie al mejor de 3 partidos, la final se juega al mejor de 5, y los porcentajes que cada equipo tenía de ganarla, previo al encuentro propiamente dicho, distinguiendo con colores los valores del modelo desarrollado por nosotros en celeste y el extraído del paper de Song, en verde.

En la llave se puede notar como el modelo de Song cambiar mucho sus valores entre fases, San Lorenzo tiene un 25.55% de pasar la primera ronda, pero luego ya es muy favorito de llevarse las 2 siguientes. Esto se debe a que diferencia más los resultados según el rival y localía y a Instituto tenía un buen historial jugando de visitante contra ellos.
Al igual que en las llaves del torneo NBA ambos modelos predijeron cosas similares y acertaron en su mayoría.

En las siguientes tablas comparamos los partidos acertados de ambos modelos, diferenciando los juegos según la probabilidad que le brindaba al equipo favorito.

## 280777

| Porcentaje 280777 | Cantidad de partidos | Acertados | Esperanza | Aciertos |
|---|---|---|---|---|
| Mas de 75% | 16 | 13 | 78.17% | 81.25% |
| Entre 65% y 75% | 90 | 65 | 69.16% | 72.22% |
| Entre 60 y 65% | 67 | 41 | 62.18% | 61.19% |
| Entre 55% y 60% | 103 | 61 | 57.39% | 59.22% |
| Entre 50% y 55% | 97 | 50 | 52.49% | 51.55% |
| TOTAL | 373 | 230 | 60.71% | 61.66% |

## Normal bivariada (Song)

| Porcentaje Song | Cantidad de partidos | Acertados | Esperanza | Aciertos |
|---|---|---|---|---|
| Mas de 75% | 243 | 171 | 89.52% | 70.37% |
| Entre 65% y 75% | 45 | 19 | 69.85% | 42.22% |
| Entre 60 y 65% | 30 | 15 | 62.87% | 50% |
| Entre 55% y 60% | 30 | 16 | 57.36% | 53.33% |
| Entre 50% y 55% | 25 | 9 | 52.74% | 36% |
| TOTAL | 373 | 230 | 79.95% | 61.66% |

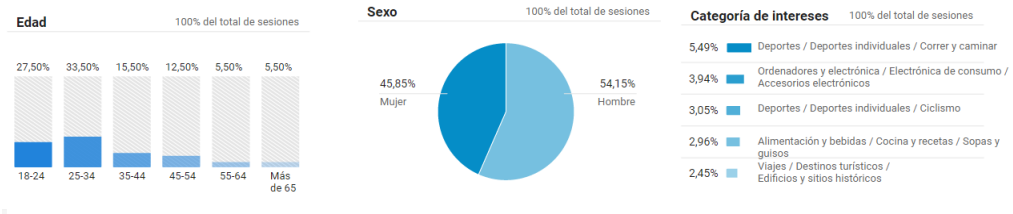
Nuevamente podemos notar como el modelo de 280777 esta más adecuado a las esperanzas de cada franja porcentual, pero comparten la misma cantidad de partidos acertados totales y, como ventaja para el algoritmo de Song, vemos un 70% de aciertos en un total de 243 partidos a los que le propició un porcentaje mayor del 75%. Si nuestro objetivo fuera acertar la mayor cantidad de partidos, sin la necesidad de apostar a todos, este método parece ser muy redituable.
Por último, comparando los modelos partido a partido podemos notar nuevamente una gran paridad. En un total de 373 partidos en 29 acertó el modelo de 28077 y no el de Song, en 29 fue al revés, en 201 los dos modelos predijeron correctamente al ganador del partido y en 114 la predicción fue errónea en ambos algoritmos.

# 5 Divulgación

Un claro objetivo del proyecto 280777 fue el de la divulgación matemática, debido a que mediante un evento mundialmente popular como un mundial de básquet y algo tan atractivo y ”mágico” como las predicciones, se puede mostrar al mundo el potencial de las matemáticas aplicadas y en particular el trabajo de nuestra facultad.
Ayudándonos con las estadísticas que nos brinda ”Google Analytics” podremos sacar algunas conclusiones sobre los usuarios de la página.

| País | Usuarios ↓ | Usuarios nuevos | Sesiones | Porcentaje de rebote | Páginas/sesión | Duración media de la sesión |
|---|---|---|---|---|---|---|
| | 539<br>% del total: 100,00 % (539) | 524<br>% del total: 100,00 % (524) | 971<br>% del total: 100,00 % (971) | 53,76 %<br>Media de la vista: 53,76 % (0,00 %) | 3,12<br>Media de la vista: 3,12 (0,00 %) | 00:04:19<br>Media de la vista: 00:04:19 (0,00 %) |
| 1. Argentina | **417** (76,80 %) | 400 (76,34 %) | 804 (82,80 %) | 49,75 % | 3,45 | 00:05:02 |
| 2. Greece | **45** (8,29 %) | 45 (8,59 %) | 49 (5,05 %) | 87,76 % | 1,18 | 00:00:16 |
| 3. Spain | **35** (6,45 %) | 35 (6,68 %) | 55 (5,66 %) | 72,73 % | 1,73 | 00:01:18 |
| 4. Chile | **9** (1,66 %) | 9 (1,72 %) | 17 (1,75 %) | 41,18 % | 2,12 | 00:02:17 |
| 5. United States | **5** (0,92 %) | 5 (0,95 %) | 8 (0,82 %) | 75,00 % | 1,25 | 00:00:21 |
| 6. Colombia | **4** (0,74 %) | 3 (0,57 %) | 5 (0,51 %) | 60,00 % | 1,60 | 00:00:35 |
| 7. Mexico | **4** (0,74 %) | 4 (0,76 %) | 4 (0,41 %) | 75,00 % | 1,25 | 00:00:07 |
| 8. Germany | **3** (0,55 %) | 3 (0,57 %) | 3 (0,31 %) | 66,67 % | 1,67 | 00:00:17 |
| 9. United Kingdom | **3** (0,55 %) | 3 (0,57 %) | 3 (0,31 %) | 100,00 % | 1,00 | 00:00:00 |
| 10. Norway | **3** (0,55 %) | 3 (0,57 %) | 5 (0,51 %) | 80,00 % | 1,40 | 00:00:21 |
| 11. Uruguay | **3** (0,55 %) | 3 (0,57 %) | 3 (0,31 %) | 33,33 % | 1,67 | 00:00:23 |
| 12. Brazil | **2** (0,37 %) | 2 (0,38 %) | 2 (0,21 %) | 100,00 % | 1,00 | 00:00:00 |
| 13. Canada | **2** (0,37 %) | 2 (0,38 %) | 2 (0,21 %) | 50,00 % | 1,50 | 00:00:56 |
| 14. France | **2** (0,37 %) | 2 (0,38 %) | 2 (0,21 %) | 50,00 % | 2,00 | 00:00:33 |
| 15. Italy | **2** (0,37 %) | 2 (0,38 %) | 2 (0,21 %) | 100,00 % | 1,00 | 00:00:00 |
| 16. Israel | **1** (0,18 %) | 1 (0,19 %) | 3 (0,31 %) | 33,33 % | 2,00 | 00:00:19 |
| 17. India | **1** (0,18 %) | 1 (0,19 %) | 1 (0,10 %) | 100,00 % | 1,00 | 00:00:00 |
| 18. Peru | **1** (0,18 %) | 0 (0,00 %) | 1 (0,10 %) | 100,00 % | 1,00 | 00:00:00 |

Cómo se ve en el cuadro anterior la página llegó a más de 500 personas de las cuales más de 400 fueron argentinos pero, curiosamente en el segundo puesto de países con más usuarios se encuentra Grecia con 45, luego más esperable España y Chile.




Analizando otras características de los usuarios, podemos ver que la gran mayoría fueron adolescentes (18-24 años) y adultos jóvenes (25-34 años).
En tema de género fue bastante parecido con una leve mayoría del lado de los hombres.

| Categoría de dispositivo | Adquisición | | | Comportamiento | | |
|---|---|---|---|---|---|---|
| | Usuarios ↓ | Usuarios nuevos | Sesiones | Porcentaje de rebote | Páginas/sesión | Duración media de la sesión |
| | 539<br>% del total: 100,00 % (539) | 524<br>% del total: 100,00 % (524) | 971<br>% del total: 100,00 % (971) | 53,76 %<br>Media de la vista: 53,76 % (0,00 %) | 3,12<br>Media de la vista: 3,12 (0,00 %) | 00:04:19<br>Media de la vista: 00:04:19 (0,00 %) |
| 1. mobile | **313** (58,07 %) | 305 (58,21 %) | 569 (58,60 %) | 63,62 % | 2,02 | 00:02:34 |
| 2. desktop | **216** (40,07 %) | 209 (39,89 %) | 391 (40,27 %) | 39,64 % | 4,74 | 00:06:57 |
| 3. tablet | **10** (1,86 %) | 10 (1,91 %) | 11 (1,13 %) | 45,45 % | 2,27 | 00:01:11 |

En las tabla anterior se puede ver que la gente que accedió al sitio web principalmente desde su celular, con sesiones cortas y explorando pocas páginas, probablemente ver la página principal, la cual contenía un resumen con las probabilidades más relevantes (candidatos a campeón, primeros de grupos y en el mundial todo lo relevante a Argentina).
A diferencia, las personas que ingresaron desde su computadora tienen un promedio de tiempo en nuestro sitio mayor con hasta el doble de páginas visitadas.

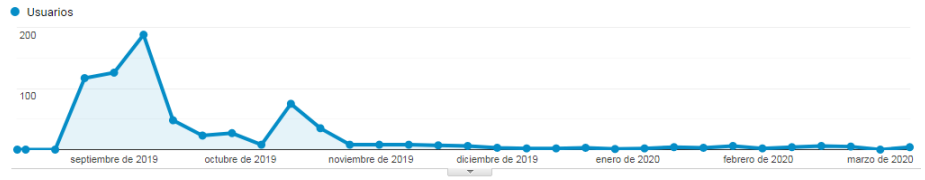
| Número de sesiones | Sesiones | Número de visitas a páginas |
|---|---|---|
| 1 | 524 | 1.053 |
| 2 | 117 | 281 |
| 3 | 50 | 167 |
| 4 | 30 | 108 |
| 5 | 23 | 51 |
| 6 | 15 | 28 |
| 7 | 10 | 32 |
| 8 | 9 | 72 |
| 9-14 | 38 | 148 |
| 15-25 | 45 | 421 |
| 26-50 | 60 | 472 |
| 51-100 | 50 | 192 |

Por último tenemos un conteo de la gente que ingresó tan sólo una vez al sitio, la gran mayoría, pero vale destacar que más de 200 personas accedieron a nuestro portal de internet en 8 o más ocasiones.

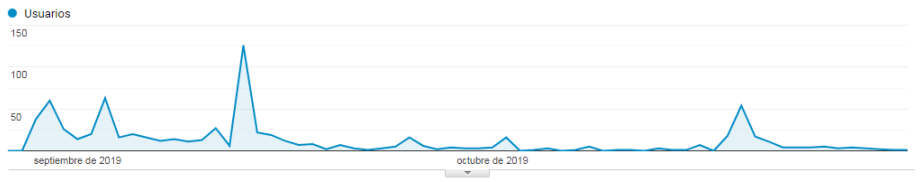
Ahora veamos la línea temporal de las visitas a nuestra página.

En el siguiente gráfico puede verse cómo el sitio web no fue muy popular para torneo local, en cuanto terminó el mundial a fines de septiembre, las visitas cayeron. Esto puede deberse en parte a que el básquet en Argentina no tiene tanta popularidad como el fútbol.


Al ver esta línea temporal un poco más detallada puede verse un pico en mitad de octubre debido a las nota periodística promocionando la página luego del comienzo de la Liga Nacional el 18 de octubre de 2019.




Por último viendo todavía más de cerca en las primeras fechas, se pueden ver unos picos en los partidos jugados por Argentina, principalmente en el partido inaugural (31/08), el último partido del grupo (4/9) y la final (15/9).



# 6 Conclusiones y Trabajos futuros

- En este trabajo se desarrolló un modelo para predecir partidos de básquet adaptando el modelo clásico de Dixon-Coles y fue probado en varios torneos de la liga nacional, el mundial de China 2019 y la temporada 2020-2021 de la NBA obteniendo buenos resultados.
- Luego de compararlo con un modelo ya existente pudimos notar que obtiene resultados muy parecidos utilizando una menor cantidad de datos, que para un deporte no tan popularizado en Argentina son muy difíciles de conseguir.
- Otra gran cualidad de nuestro modelo es la velocidad para obtener las probabilidades de los partidos debido a su simpleza, se basa en una regresión lineal generalizada, por lo que se pueden ajustar rápido estas predicciones en cuanto van terminando los partidos de un torneo.
- En esta tesis también pudimos ver cómo los argentinos nos enloquecemos por la selección, los picos de gente que ingresó a la página en el mundial, esos momentos deportivos se pueden aprovechar para divulgar la matemática aplicada mediante este tipo trabajos en disciplinas que, a primera vista, están muy alejadas.
- Nos queda como trabajo a futuro aplicarlo para otras ligas, principalmente las europeas, que son las más importantes después de la NBA.
- Otro modelo para agregar es uno que pueda calcular los puntajes en el tiempo extra de un partido, el cual puede ayudar a predecir mejor a los partidos que son muy parejos y se espera que vayan a esta instancia del juego.
- Cómo último se podría complementar el modelo con uno que incluya datos sobre las formaciones de los equipos, lesiones o el cansancio de los jugadores debido a la sobrecarga de partidos lo que puede llevar a algunos resultados impensables si ambos estuvieran en óptimas condiciones.

# BIBLIOGRAFIA

[KS06]: Kvam P. & Sokol J. ”A logistic regression/Markov chain model for NCAA basketball” (2006), Naval Research Logistics, Vol. 53 Nº8, pag. 788-803.

[SZS18]: Song K., Zou Q. & Shi J.”Modelling the scores and performance statistics of NBA basketball games” (2018), Communications in Statistics - Simulation and Computation, Volumen 49, Pag. 2604-2616.

[HJLL23]: Horvat, T., Job, J., Logozar, R. & Livada, Č., ”A Data-Driven Machine Learning Algorithm for Predicting the Outcomes of NBA Games” (2023). Symmetry, Vol. 15 Nº 4, Art. 798.

[LBB22]: Loeffelholz B., Bednar E. & Kenneth W Bauer”Predicting NBA Games Using Neural Networks” (2022), Journal of Quantitative Analysis in Sports, Vol. 5 Nº 1, Art. 7

[CSF21]:Cohan A., Schuster J. & Fernandez J., ”A deep learning approach to injury forecasting in NBA basketball” (2021). Journal of Sports Analytics, Vol. 7 Nº 4, pag. 277-289

[DC97]: Dixon M. & Coles S.,”Modelling association football scores and inefficiencies in the football betting market” (1997). Journal of the Royal Statistical Society Series C (Applied Statistics), Vol. 46, Nº 2, pag. 265-280.