



# UNIVERSIDAD TORCUATO DI TELLA

Cómo encontrar el mejor jugador para tu Equipo de Fútbol

Escuela de Negocios - Licenciatura en Tecnología Digital

Tomás Glauberman\*

Ignacio Pardo†

Juan Ignacio Silvestri‡

**CABA, Argentina. Diciembre 2024**

## Abstract

En la última década, el análisis deportivo ha evolucionado hacia una perspectiva cada vez más matemática y sofisticada. Aplicaciones como el uso de análisis espacial en Basketball (Goldsberry, 2012) y modelos de juegos de suma cero en fútbol (Hirotsu y Wright, 2006) son ejemplos claros de la tendencia creciente en este campo. El béisbol, por mucho tiempo el deporte preferido para la analítica, ha experimentado una profunda transformación con la implementación de Sabermetrics (Baumer y Zimbalist, 2014; Wolf, 2015). La introducción de herramientas analíticas avanzadas ha producido resultados positivos para muchos equipos, lo que subraya el valor de estudiar métricas específicas dentro de cada deporte.

Este desarrollo se centra en el fútbol, un deporte en el cual los análisis previos se han concentrado, en su mayoría, en predecir resultados de partidos y mejorar el rendimiento de los equipos. Sin embargo, este trabajo propone un enfoque diferente al analizar el impacto de los jugadores sobre la posesión de balón y los disparos del equipo desde una perspectiva probabilística.

A partir de la métrica PSL (Huang et al., n.d.), planteamos un proceso para comparar el impacto que tienen los jugadores sobre la performance del equipo. En un enfoque bayesiano, logramos formular una metodología para estudiar la distribución de la performance de un equipo. Además, desarrollamos un modelo de machine learning llamado Player2Vec sobre el modelo teórico de redes de jugadores para representar a cada jugador. De esta manera podemos hallar jugadores similares para luego comparar su rendimiento resultante en un nuevo equipo.

---

\*21F78 | tglauberman@mail.utdt.edu

†21R1160 | ipardo@mail.utdt.edu

‡21Q111 | jsilvestri@mail.utdt.edu

# Índice

0.1	Agradecimientos . . . . .	3
0.2	Introducción . . . . .	3
0.3	Motivación Justificación del tema . . . . .	3
0.3.1	Relevancia Académica . . . . .	3
0.3.2	Relevancia Práctica . . . . .	3
0.4	Objetivos de Proyecto . . . . .	3
0.4.1	Objetivo General . . . . .	3
0.4.2	Objetivos Específicos . . . . .	3
0.5	Definición del problema . . . . .	5
0.5.1	PSL como métrica de Performance . . . . .	5
0.5.2	Modelo de Red de Jugadores . . . . .	5
0.5.3	Modelo Predictivo de probabilidades de transición . . . . .	6
0.5.4	Test de Sensibilidad sobre PSL . . . . .	8
0.5.5	Modelo Predictivo sobre $r(J, S)$ . . . . .	8
0.6	Analisis de las distribuciones de los $r(J, S)$ . . . . .	9
0.6.1	Comparación de las distribuciones de los $r(J, S)$ . . . . .	9
0.7	Estimación de la Distribución del PSL . . . . .	14
0.7.1	Variables Aleatorias para los $r(U, V)$ y PSL por Priors . . . . .	14
0.7.2	Proceso de Monte Carlo para estimar la distribución del PSL . . . . .	14
0.7.3	Comparar el impacto sobre el PSL de dos jugadores en una formación . . . . .	16
0.8	Player2Vec: Embeddings de Jugadores . . . . .	16
0.8.1	Modelado de la EPL 2012/13 como Grafo . . . . .	17
0.8.2	Implementación . . . . .	18
0.8.3	Visualización y Exploración de los Embeddings . . . . .	20
0.8.4	Potencial de Player2Vec . . . . .	23
0.9	Hipótesis . . . . .	24
0.10	Marco teórico . . . . .	24
0.11	Marco metodológico . . . . .	24
0.12	Resultados . . . . .	24
0.13	Discusión . . . . .	24
0.14	Conclusiones & Recomendaciones {#conclusiones-&-recomendaciones} . . . . .	25
0.15	Referencias bibliográficas . . . . .	26
0.16	Apéndices: Tablas, figuras, anexos {#apéndices:-tablas,-figuras,-anexos} . . . . .	26

# Índice

## 0.1 Agradecimientos

Este trabajo no hubiera sido posible sin la ayuda de los profesores Gustavo Vulcano (Escuela de Negocios, Universidad Torcuato Di Tella) y Santiago Gallino (The Wharton School, University of Pennsylvania). Además queremos agradecer a Ignacio Vigilante (TIC - Escuela ORT) y Tomás Spognardi (Exactas - UBA) por sus aportes al modelo de Player2Vec y al PSL Bayesiano respectivamente.

## 0.2 Introducción

Presenta el tema del trabajo, su contexto y la importancia. Aquí se debe captar el interés del lector, explicando brevemente los aspectos más importantes que se desarrollarán.

## 0.3 Motivación Justificación del tema

Explica **por qué el tema elegido es relevante**, tanto a nivel académico como práctico. Debe argumentar la importancia del trabajo para el campo de estudio o la sociedad. Similar a lo que se completó en el formulario de licitación de proyectos iniciales.

El fútbol es uno de los deportes más populares y seguidos en todo el mundo. La capacidad de un equipo para ganar partidos y campeonatos depende en gran medida de la calidad y el rendimiento de sus jugadores. En este contexto, la identificación y selección de los mejores jugadores para un equipo se convierte en una tarea crucial para entrenadores, directores deportivos y analistas de rendimiento.

### 0.3.1 Relevancia Académica

Desde una perspectiva académica, el análisis del rendimiento de los jugadores de fútbol ha sido un área de interés creciente en los últimos años. La aplicación de técnicas avanzadas de análisis de datos, aprendizaje automático y modelos probabilísticos ha permitido una comprensión más profunda del impacto de los jugadores en el rendimiento del equipo. Algunos ejemplos del estado del arte incluyen el modelo para maximizar la posesión esperada propuesto en el artículo de Rahimian et al. (2023) (Rahimian et al., 2023) y el modelo de redes de jugadores para calcular la probabilidad de disparar al arco antes de perder el balón (PSL) presentado en el trabajo de Huang et al.(Huang et al., n.d.).

Este trabajo se enmarca en esta línea de investigación, contribuyendo al desarrollo de nuevas metodologías y herramientas para evaluar y comparar el rendimiento de los jugadores.

### 0.3.2 Relevancia Práctica

En el ámbito práctico, la capacidad de identificar a los mejores jugadores tiene implicaciones directas en la toma de decisiones estratégicas y operativas de los equipos de fútbol. La correcta selección de jugadores puede mejorar significativamente el rendimiento del equipo, aumentar las probabilidades de éxito en competiciones y optimizar la inversión en fichajes. Además, el uso de modelos avanzados como Player2Vec y PSL Bayesiano proporciona una ventaja competitiva por su poder predictivo del rendimiento de los jugadores.

## 0.4 Objetivos de Proyecto

### 0.4.1 Objetivo General

El objetivo principal de este proyecto es desarrollar y aplicar modelos avanzados de análisis de datos y probabilísticos, para mejorar la evaluación, comparación y selección de jugadores de fútbol. Esto permitirá a los equipos tomar decisiones más informadas y estratégicas, optimizando su rendimiento y aumentando sus probabilidades de éxito en competiciones. Mas concretamente, este trabajo busca responder la pregunta del título “¿Cómo encontrar al jugador ideal para tu equipo de fútbol?”.

### 0.4.2 Objetivos Específicos

#### 1. Desarrollar un Modelo de Evaluación del Rendimiento de Jugadores:

- Implementar el modelo Player2Vec para analizar y representar las transiciones entre estados de los jugadores.

- Utilizar el modelo PSL Bayesiano para estimar el impacto de los jugadores en el rendimiento del equipo.

**2. Comparar el Rendimiento de Jugadores:**

- Establecer métricas estandarizadas para comparar objetivamente el rendimiento de jugadores en diferentes posiciones y roles.
- Aplicar técnicas de aprendizaje automático y análisis de datos para identificar patrones y tendencias en el rendimiento de los jugadores.

**3. Optimizar la Selección de Jugadores:**

- Desarrollar un sistema de recomendación para identificar a los jugadores que mejor se adaptan a las necesidades y estrategias específicas de un equipo.
- Evaluar la efectividad del sistema de recomendación mediante estudios de caso y análisis de datos históricos.

**4. Validar los Modelos:**

- Realizar pruebas y validaciones de los modelos desarrollados utilizando datos reales de partidos y jugadores.

**5. Generar Conocimiento y Herramientas para la Comunidad:**

- Documentar y publicar los resultados y metodologías desarrolladas en el proyecto.
- Crear herramientas y recursos accesibles para entrenadores, analistas y directores deportivos que deseen aplicar estos modelos en sus equipos.

## 0.5 Definición del problema

A partir de la pregunta de la investigación, se plantea el problema de encontrar el jugador ideal para un equipo de fútbol. En un comienzo nos encontramos planteando como definir *performance* de un jugador y cómo compararla con otros jugadores. Surgió la necesidad de encontrar una métrica evaluar el impacto de un jugador en el rendimiento de un equipo y como definir estos agentes. Además es necesario poder representar concretamente a un Jugador  $J$ .

### 0.5.1 PSL como métrica de Performance

En el paper en proceso *How to Find the Right Player for your Soccer Team?* (Huang et al.) se plantea la descomposición del Gol Esperado ( $xG$ ) como:

$$xG(A) = P(A) \cdot PSL(A) \cdot SA(A)$$

Donde  $A$  es el equipo,  $P(A)$  es la posesión del balón,  $PSL(A)$  es la probabilidad patear al arco antes de perder el balón y  $SA(A)$  es la probabilidad de que un disparo al arco se convierta en gol. A diferencia de la posesión del balón y la probabilidad de convertir un disparo en gol,  $PSL(A)$  no es una métrica comúnmente utilizada en el análisis de fútbol ni existen modelos que la calculen. El paper plantea un modelo de red de jugadores que permite calcular  $PSL(A)$  para cada equipo.

### 0.5.2 Modelo de Red de Jugadores

Utilizando Cadenas de Markov de Tiempo Continuo (CTMC) se puede calcular la probabilidad de que un equipo pierda el balón antes de patear al arco. En este modelo de red de jugadores se plantea un modelo de 14 estados: 11 jugadores ( $J_1 \dots J_{11}$ ), Ganancia, Pérdida y Disparo.

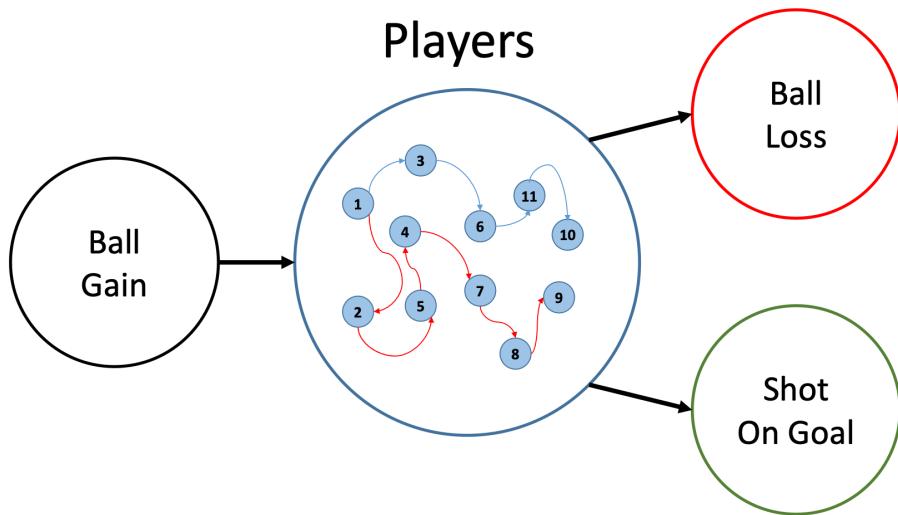


Figure 1: Modelo de Red de Jugadores

El grafo presentado en la figura representa el modelo de red de jugadores. Cada nodo representa un estado y cada arista representa una transición entre estados. El nodo verde representa el estado de disparo al arco, el rojo la pérdida del balón y el azul la ganancia del balón por parte de un jugador. Los ejes entre los nodos se representan con una matriz de adyacencia  $R$  donde cada valor  $r(U, V)$  representa la ratio de transición entre los estados  $U$  y  $V$ .

$$R = \begin{pmatrix} 0 & r(G, J_1) & \dots & r(G, J_{11}) & 0 & 0 \\ 0 & 0 & \dots & r(J_1, J_{11}) & r(J_1, L) & r(J_1, S) \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & r(J_{11}, J_1) & \dots & 0 & r(J_{11}, L) & r(J_{11}, S) \\ 0 & 0 & \dots & 0 & 1 & 0 \\ 0 & 0 & \dots & 0 & 0 & 1 \end{pmatrix}$$

A partir de la matriz de ratio de acción sobre tiempo jugado  $R$  (ganancias, pases a otro jugador, disparos o pérdidas) se puede obtener la matriz de transición de estados  $Q$  para el CMTTC de normalizar las filas de  $R$ :

Para cada par de estados  $U$  y  $V$  se define  $q(U, V) = \frac{r(U, V)}{\sum_{i=1}^{14} r(U, i)}$

$$Q = \begin{pmatrix} 0 & q(G, J_1) & \dots & q(G, J_{11}) & 0 & 0 \\ 0 & 0 & \dots & q(J_1, J_{11}) & q(J_1, L) & q(J_1, S) \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & q(J_{11}, J_1) & \dots & 0 & q(J_{11}, L) & q(J_{11}, S) \\ 0 & 0 & \dots & 0 & 1 & 0 \\ 0 & 0 & \dots & 0 & 0 & 1 \end{pmatrix}$$

Finalmente a partir de la matriz de probabilidades de transición  $Q$  se puede calcular  $PSL(A)$  como:

$$PSL(A) = [1, 0, \dots, 0] \cdot (I - T)^{-1} \cdot X \cdot [0, 1]^T$$

Siendo  $T$  las probabilidades de transición de los estados transitorios,  $X$  las probabilidades de transición de los estados transitorios a los estados absorbentes e  $I$  la matriz identidad.

A partir de este modelo en el paper se evaluó para una temporada de la Premier League (EPL 2012/13) la diferencia entre los PSL de cada equipo y luego de forma empírica se demuestra como el  $PSL(A)$  tiene alta correlación positiva con el rendimiento del equipo por sobre el contrincante. Finalmente hayamos una métrica significativa de rendimiento de un equipo en la métrica  $PSL$ . Sin embargo, da a lugar a la investigación de como se puede aplicar esta métrica a nivel de jugador y como se puede comparar el rendimiento de jugadores en distintos equipos.

Para evaluar el impacto de un jugador  $J$  se debe, o bien conocer la probabilidad de transición entre  $J$  y los otros 13 estados (10 jugadores, Ganancia, Pérdida y Disparo) o bien lograr estimar la probabilidad de transición entre  $J$  y los otros 13 estados.

En este trabajo se propone un método probabilístico bayesiano para hallar la Distribución del PSL dada la distribución de probabilidades de transición entre cada uno de los 11 jugadores y los otros 13 estados.

### 0.5.3 Modelo Predictivo de probabilidades de transición

En un comienzo se planteó desarrollar un modelo predictivo para estimar las ratios de transición entre los estados. Optamos por buscar predecir los ratios  $r$  y no las probabilidades de transición  $q$  ya que la normalización no es igual en cada instancia de  $R$ . Mas concretamente buscamos estimar la función  $f$  que mapea los estados  $U$  y  $V$  a la ratio de transición  $r(U, V)$ .

$$\hat{r}(U, V) = f(U, V, \theta)$$

Comenzamos armando un modelo para predecir únicamente los ratios de pases  $r(J_i, J_j)$  entre un jugador  $J_i$  y otro jugador  $J_j$ . Lo que correspondiera a los siguientes valores de la matriz  $R$ :

$$R = \begin{pmatrix} 0 & r(G, J_1) & \dots & r(G, J_{11}) & 0 & 0 \\ 0 & 0 & \dots & r(J_1, J_{11}) & r(J_1, L) & r(J_1, S) \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & r(J_{11}, J_1) & \dots & 0 & r(J_{11}, L) & r(J_{11}, S) \\ 0 & 0 & \dots & 0 & 1 & 0 \\ 0 & 0 & \dots & 0 & 0 & 1 \end{pmatrix}$$

Para poder utilizar un modelo de machine learning tradicional necesitamos de poder representar a cada jugador  $J$  de forma vectorial. Armamos un vector de métricas agregadas para un jugador al momento del partido a predecir. Estas métricas incluyen la cantidad de pases, disparos, goles, pérdidas, etc. sobre el total de tiempo jugado, ademas de el equipo en el que juega.

$$J = [\text{Passes}/90, \text{Shots}/90, \text{Goals}/90, \text{Losses}/90, \text{Time Played}, \text{Team ID}]$$

Para el modelo predictivo comenzamos utilizando un modelo de XGBoost para la regresión pero rápidamente observamos que por la naturelza de arbol al predecir con la media de las observaciones por hoja las predicciones resultaban casi discretas, por lo que viramos a explorar un modelo de regresión lineal para predecir los ratios de pases entre jugadores.

Para validar elegimos separar de forma temporal los 380 partidos de la temporada 2012/13 de la EPL: los primeros 269 partidos de entrenamiento; los últimos 111 de test ( $\mu + 2/3\sigma$ ). Ademas para construir el dataset, elegimos agarrar parejas de jugadores de los partidos de Train y removerlos de los mismos para poder en Test predecir ratios de transición entre jugadores que no se vieron en Train.

Luego de entrenar el modelo, para cada instancia de test obtuvimos la matriz de ratios de transición  $R$  y calculamos el PSL real, para luego predecir la matriz de transición  $\hat{R}$  y calcular el PSL predicho. Finalmente calculamos el coeficiente de correlación de Pearson entre el PSL real y el PSL predicho.

En el siguiente gráfico podemos observar como a pesar de predecir muy pobre los ratios de transición al resultar en un coeficiente de correlación de Pearson entre los  $r(J_i, J_j)$  y los  $\hat{r}(J_i, J_j)$  de 0.12, sin embargo al comparar el PSL real del PSL calculado a partir de  $\hat{R}$  se obtiene un coeficiente de correlación de Pearson de 0.85.

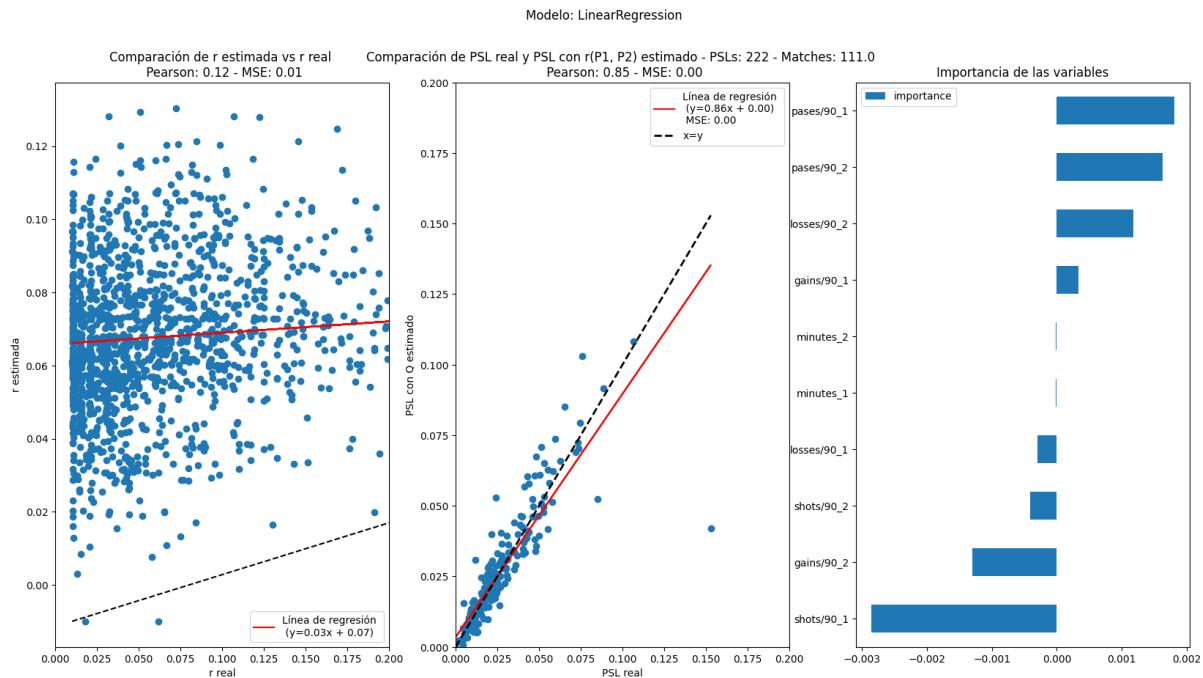


Figure 2: Resultados Modelo de Regresión Lineal

El modelo planteado no es capaz de predecir los ratios de transición, y a pesar de que desarrollamos otros modelos como XGBoost para regresión, Redes Neuronales y Redes Neuronales Probabilísticas (PNNs) no es posible predecir los ratios de transición entre los estados a partir de las métricas de los jugadores. Esto se debe principalmente a la cantidad de datos y la poca relación entre ellos. Al evaluar como resolver la predicción de los  $r(J_i, J_j)$  decidimos observar como cada ratio de transición afecta al PSL.

#### 0.5.4 Test de Sensibilidad sobre PSL

#### 0.5.5 Modelo Predictivo sobre $r(J, S)$

Luego de lo observado con el Test de Sensibilidad sobre PSL, decidimos cambiar el enfoque de la predicción de los ratios de transición entre jugadores a la predicción de los ratios de transición entre jugadores y el estado de disparo al arco. Esto se debe a que al observar la matriz de ratios de transición  $R$  se observa que los ratios de transición entre jugadores y el estado de disparo al arco son los que más afectan al PSL.

El nuevo modelo se enfoca en la siguiente sección de la matriz  $R$ :

$$R = \begin{pmatrix} 0 & r(G, J_1) & \dots & r(G, J_{11}) & 0 & 0 \\ 0 & 0 & \dots & r(J_1, J_{11}) & r(J_1, L) & r(J_1, S) \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & r(J_{11}, J_1) & \dots & 0 & r(J_{11}, L) & r(J_{11}, S) \\ 0 & 0 & \dots & 0 & 1 & 0 \\ 0 & 0 & \dots & 0 & 0 & 1 \end{pmatrix}$$

Para el vector de los jugadores  $J$  se agregó tambien la posición en la que juega (Arquero, Defensor, Mediocampista, Delantero) one-hot-encoded.

Luego se entrenó un modelo de XGBoost para Regresión con el mismo split de Train y Test. Se logró obtener un mejor resultado sobre la predicciones de Train en comparación al modelo anterior, sin embargo al evaluar en Test. Se obtuvo un coeficiente de correlación de Pearson de 0.95 entre los  $r(J_i, S)$  y los  $\hat{r}(J_i, S)$  en Train, pero de 0.08 en Test.

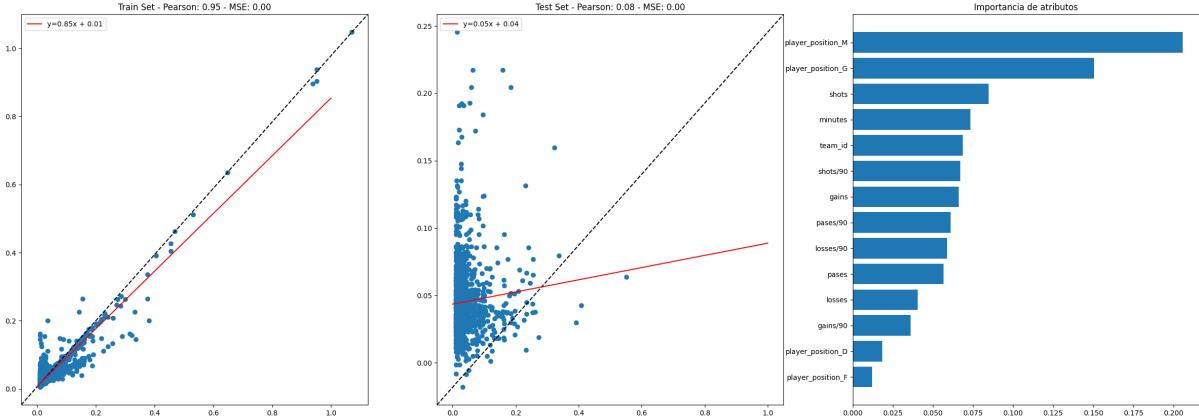


Figure 3: Resultados Modelo de XGBoost

Este resultado junto al del modelo de predicción de ratios de pases nos llevó a buscar una mejor representación vectorial de los jugadores. En la sección de Player2Vec se explica el modelo utilizado para obtener un vector de representación (embedding  $E$ ) de cada jugador. Con este embedding por construimos una red neuronal, el modelo resultante  $f(E_J, \text{partido})$  dado el embedding de los jugadores y el partido predice los ratios de transición entre jugadores y el estado de disparo al arco.

## 0.6 Análisis de las distribuciones de los $r(J, S)$

En un esfuerzo de comprender mejor el modelo de ratios de transición entre jugadores y el estado de disparo al arco, se decidió analizar las distribuciones de los  $r(J, S)$  para cada jugador en la temporada 2012/13 de la EPL.

Se observó que las distribuciones de los ratios de transición entre jugadores y el estado de disparo tienen moda cercana a 0, lo que indica que la mayoría de los jugadores tienen una baja probabilidad de disparar al arco antes de perder el balón. En la siguiente figura se puede observar la distribución de los  $r(J, S)$  para todos los jugadores de la temporada 2012/13 de la EPL en todos los partidos.

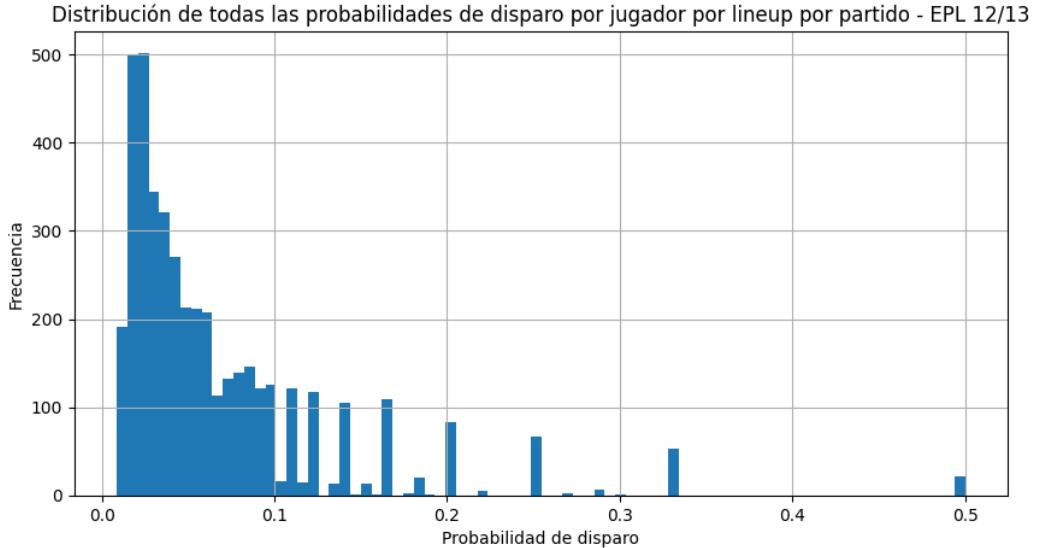


Figure 4: Distribución de todos los  $r(J, S)$

Además, se observó que la distribución de los  $r(J, S)$  de cada jugador no necesariamente sigue una distribución normal ni similar a la de otros jugadores.

Para el siguiente análisis se ajustaron las distribuciones de los  $r(J, S)$  de cada jugador a una distribución de probabilidad beta y se obtuvieron los parámetros  $\alpha$  y  $\beta$  de cada jugador.

Inicialmente presentamos la distribución de dos jugadores a modo de ejemplo: **Sergio Agüero** y **Robin van Persie**

Luego se analizó la distribución de los  $r(J, S)$  de los 10 jugadores con mayor cantidad de disparos, con mayor sesgo y con mayor suma de disparos a modo de comparación.

### 0.6.1 Comparación de las distribuciones de los $r(J, S)$

A partir de la distribución ajustada de un jugador, podemos luego hayar por ejemplo jugadores similares en base a la distribución de los  $r(J, S)$  utilizando la divergencia de Kullback-Leibler (KL).

En el siguiente gráfico se observa la distribución de los  $r(J, S)$  de jugadores similares a él en la temporada 2012/13 de la EPL. Además se presentan solapados en otra figura.

Finalmente podemos agregar la condición de *misma posición* al comparar dos jugadores, en el caso de Agüero de Delantero (F por Forward) y hayar nuevamente jugadores aún más similares a él.

Para conocer mejor la varianza de las distribuciones de los  $r(J, S)$  de los jugadores, se estudió la distribución de los parámetros  $\alpha$  y  $\beta$  de las distribuciones beta ajustadas. Hicimos un análisis de clustering para agrupar a los jugadores en base a sus distribuciones de los  $r(J, S)$ .

Como un extra, este sistema de clustering nos permite hallar rápido jugadores similares entre sí. A partir de los clusters la siguiente figura presenta las posibles distribuciones en cada cluster.

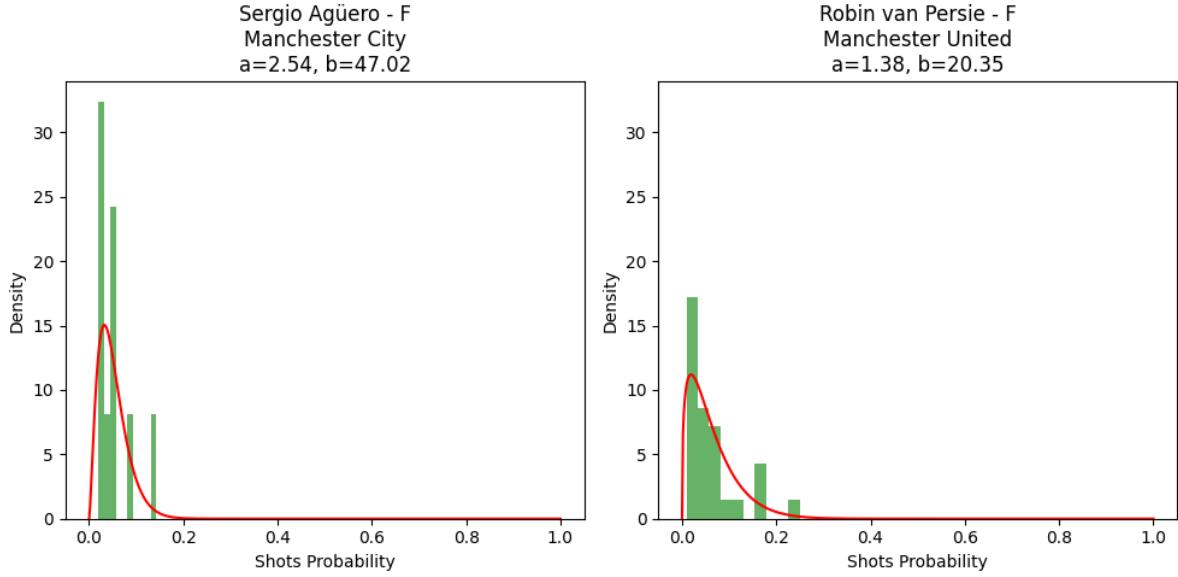


Figure 5: Distribución de los  $r(J, S)$  de Sergio Agüero y Robin van Persie

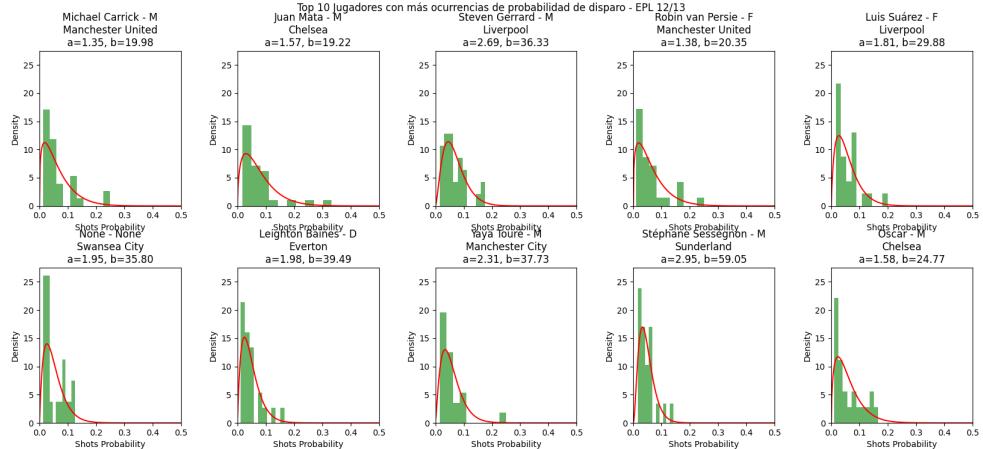


Figure 6: Distribución de los  $r(J, S)$  de los 10 jugadores con mayor cantidad de disparos

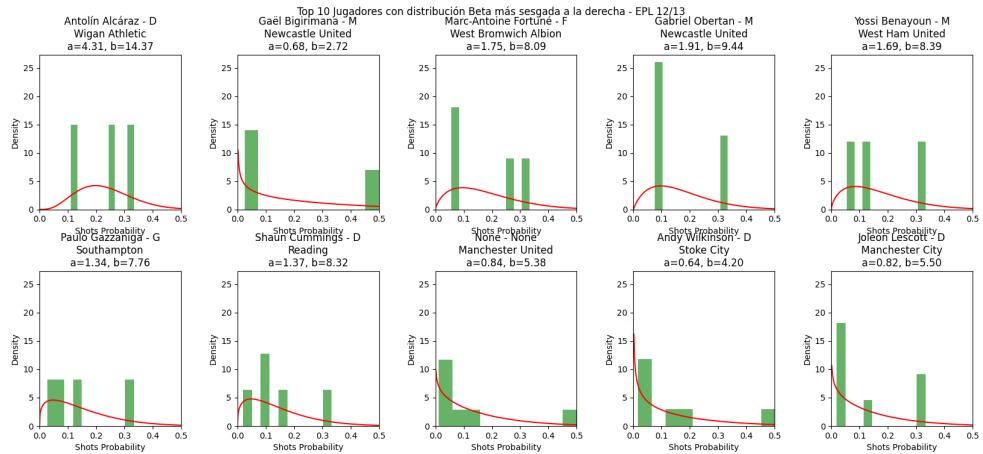


Figure 7: Distribución de los  $r(J, S)$  de los 10 jugadores con mayor sesgo

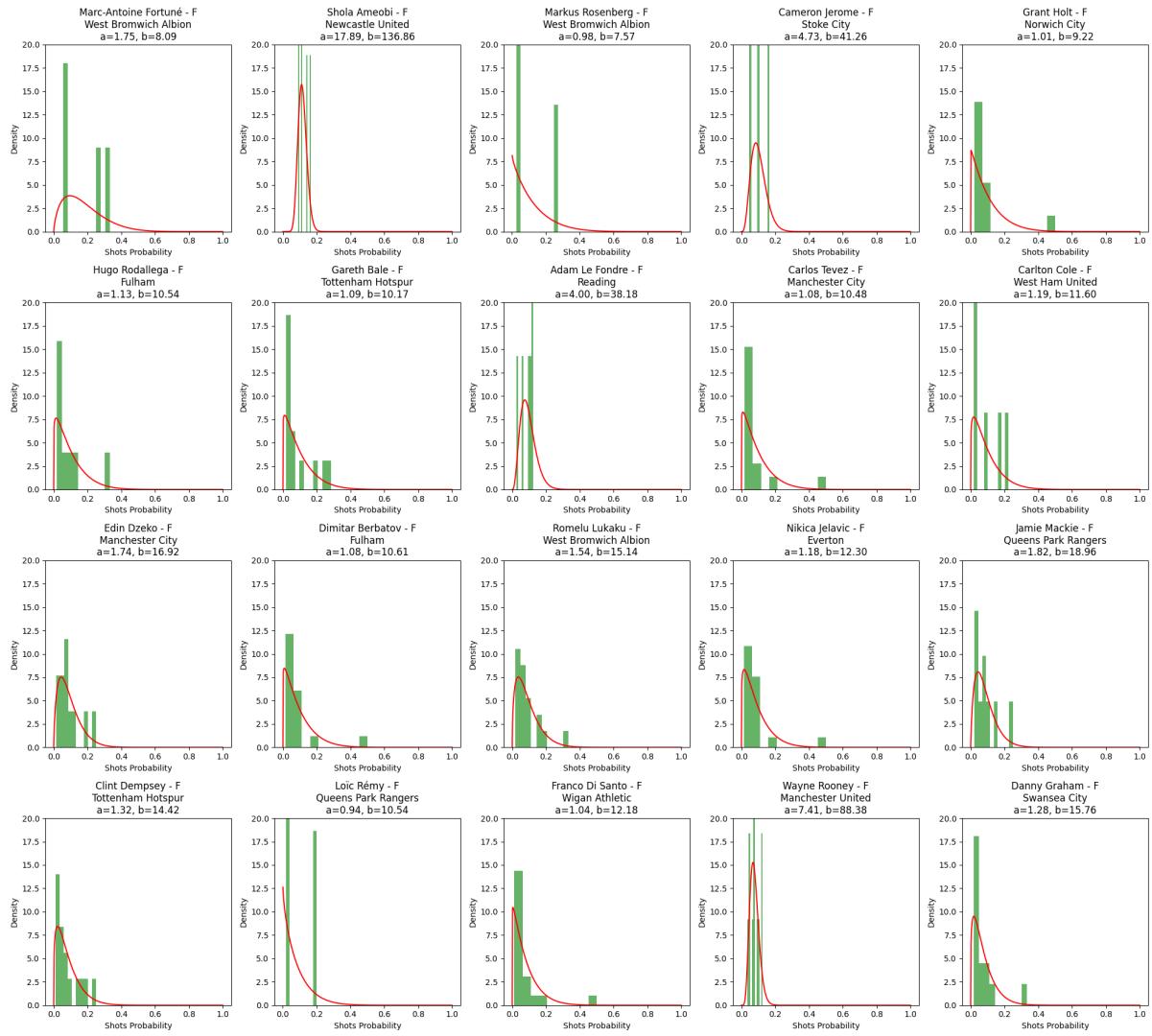


Figure 8: Top 20 Delanteros con distribución Beta más sesgada a la derecha - EPL 12/13

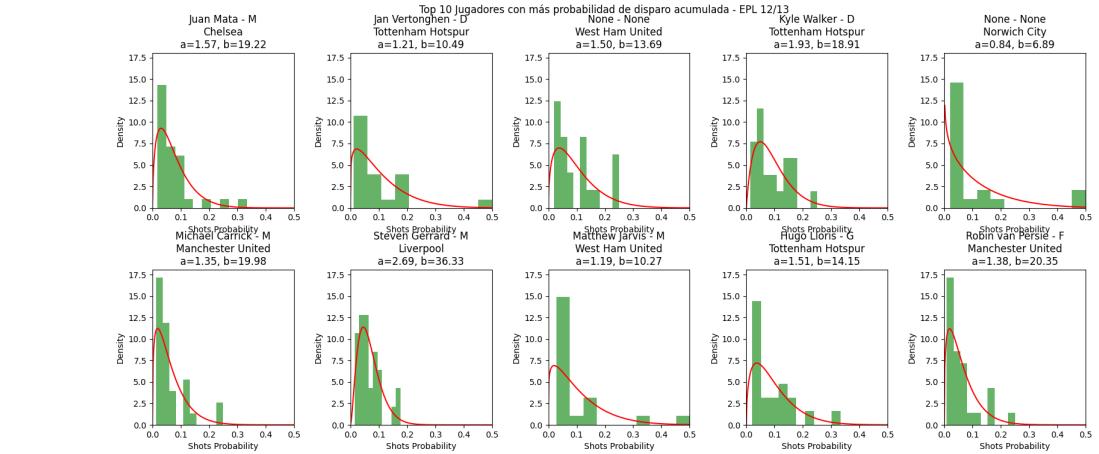


Figure 9: Distribución de los  $r(J, S)$  de los 10 jugadores con mayor suma

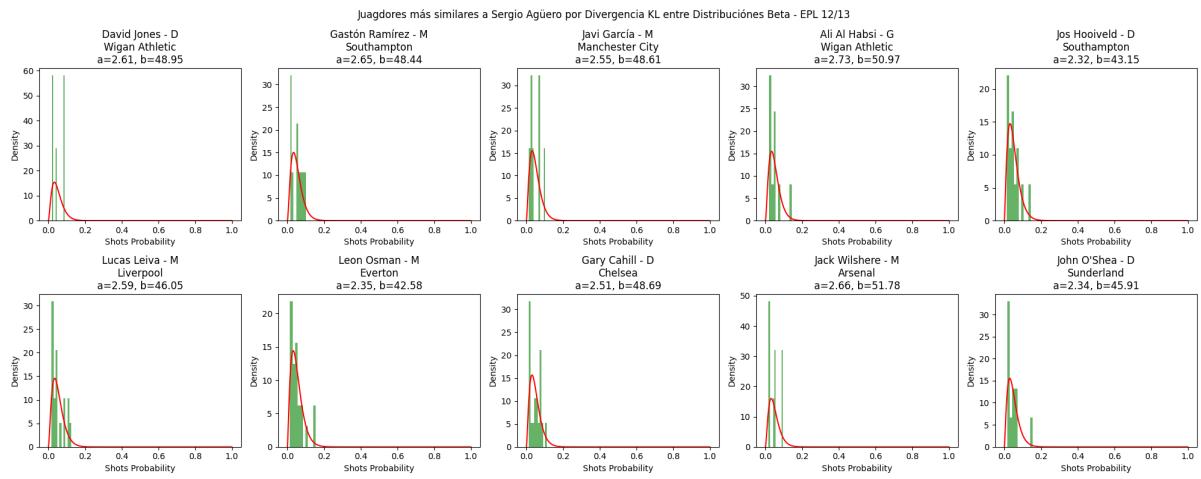


Figure 10: Distribución de los  $r(J, S)$  de jugadores similares a Sergio Agüero

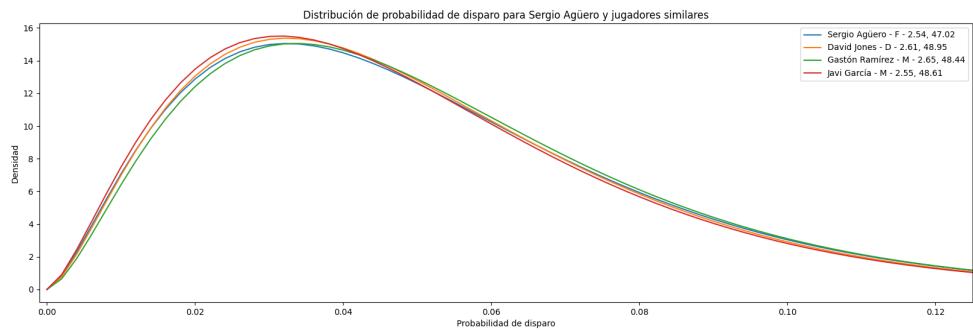


Figure 11: Distribución de los  $r(J, S)$  de jugadores similares a Sergio Agüero Superpuestos

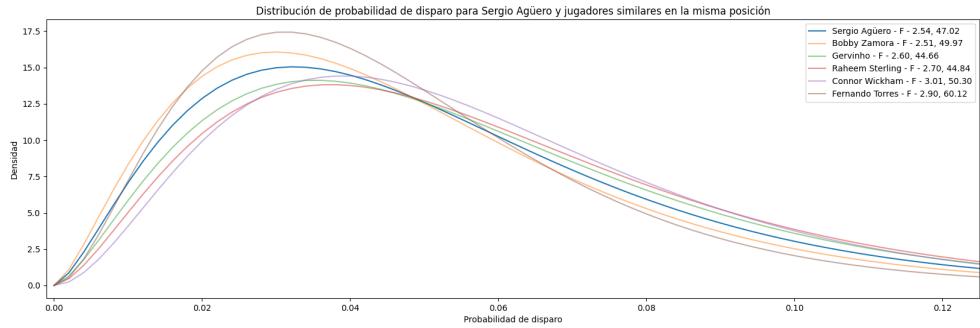


Figure 12: Distribución de los  $r(J, S)$  de jugadores similares a Sergio Agüero de la misma posición

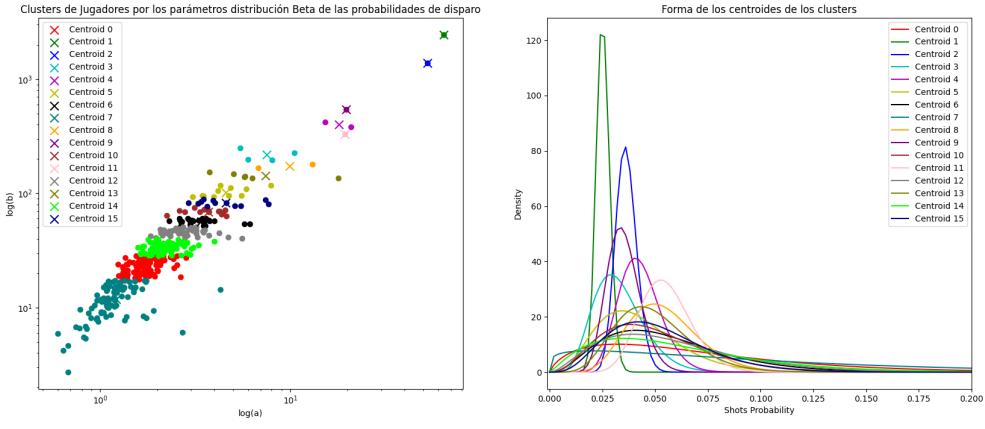


Figure 13: Distribución de los parámetros  $\alpha$  y  $\beta$  de los  $r(J, S)$  de los jugadores

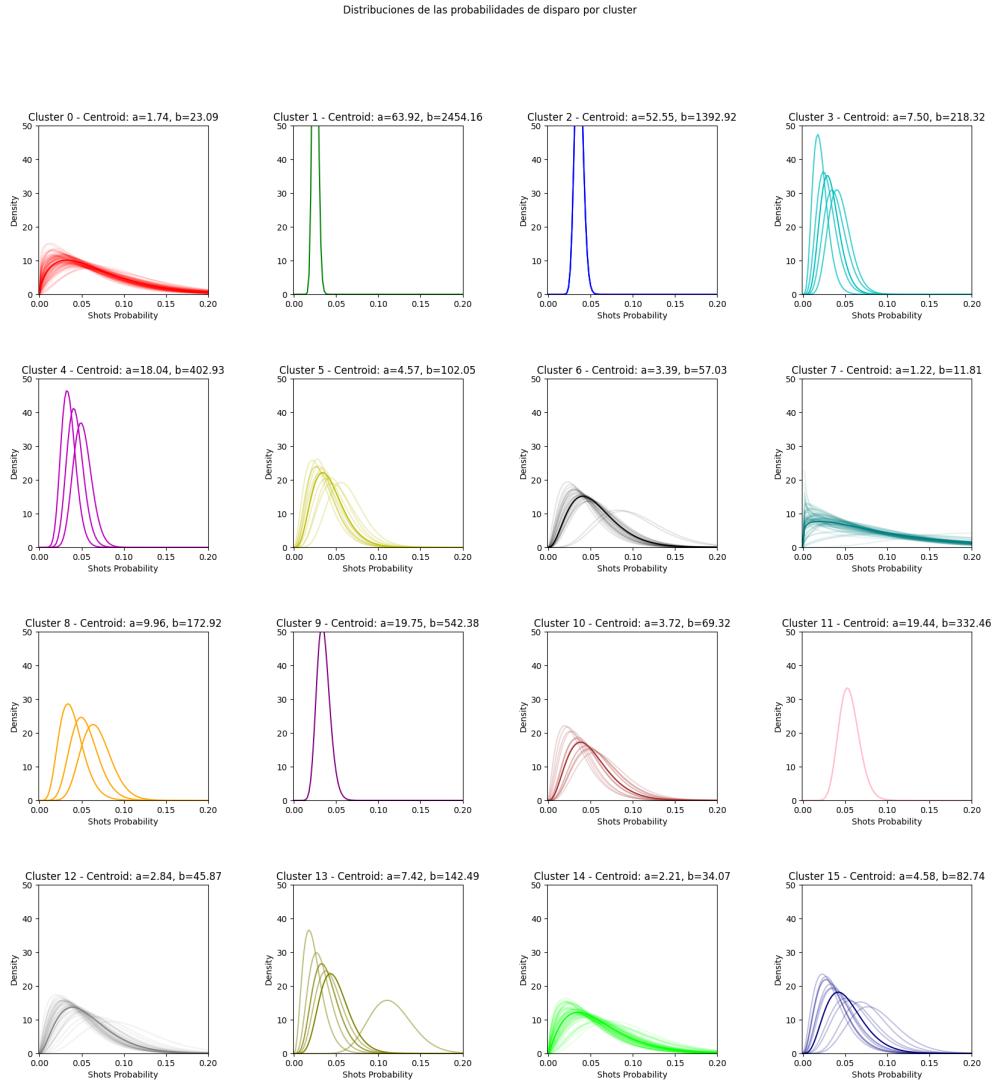


Figure 14: Distribución de los  $r(J, S)$  de jugadores en clusters

## 0.7 Estimación de la Distribución del PSL

A partir de los resultados obtenidos en el análisis de las distribuciones de los  $r(J, S)$ , se propone un utilizar estas como priors para cada jugador, es decir, se asume que la distribución de los  $r(J, S)$  de un jugador es la distribución a priori de la variable aleatoria  $r(J, S)$  para ese jugador, lo mismo para los  $r(J_i, J_j)$ , los  $r(J, L)$  y los  $r(J, G)$ .

De esta forma, cada jugador  $J$  tiene una distribución a priori para cada uno de los 14 estados, considerando esto, podemos reformular la matriz de ratios de transición como una matriz de variables aleatorias donde cada una se distribuye según la distribución a priori del jugador correspondiente.

### 0.7.1 Variables Aleatorias para los $r(U, V)$ y PSL por Priors

Para actualizar la notación, sean  $r_{J,V}$  la variable aleatoria que representa la ratio de transición entre el jugador  $J$  y el estado  $V$ , esto incluye  $r_{J,S}$ ,  $r_{J,L}$  y tambien  $r_{G,J}$ , asi como los  $r_{J_i,J_j}$  para  $i, j \in [1, 11]$ .

Luego  $r_{J,V} \sim F_x$  la distribución a priori de la variable aleatoria  $r_{J,V}$ .

Para generalizar el analisis de distribuciones planteadas en la sección anterior, se propone utilizar una distribución KDE (Kernel Density Estimation) a partir de los histogramas de los  $r(J, V)$  para modelar sus distribuciones, ya que no todos los ratios de transición siguen una distribución beta tan bien como los  $r(J, S)$ .

Finalmente obtenemos, para una formación dada de 11 jugadores, una matriz de variables aleatorias  $\mathbf{R}$ .

$$\mathbf{R} = \begin{pmatrix} 0 & r_{G,J_1} & \dots & r_{G,J_{11}} & 0 & 0 \\ 0 & 0 & \dots & r_{J_1,J_{11}} & r_{J_1,L} & r_{J_1,S} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & r_{J_{11},J_1} & \dots & 0 & r_{J_{11},L} & r_{J_{11},S} \\ 0 & 0 & \dots & 0 & 1 & 0 \\ 0 & 0 & \dots & 0 & 0 & 1 \end{pmatrix}$$

Para mejor claridad, la siguiente vizualización muestra la matriz de variables aleatorias  $\mathbf{R}$  para un equipo de ejemplo. En cada posición se observa la distribución a priori de la variable aleatoria correspondiente.

### 0.7.2 Proceso de Monte Carlo para estimar la distribución del PSL

Dado un equipo  $A$  con una formación de 11 jugadores, se busca estimar la distribución del PSL de ese equipo a partir de las distribuciones a priori de los  $r(U, V)$  de cada jugador. Para ello, se propone un proceso de Monte Carlo para muestrear de las distribuciones a priori de los  $r(U, V)$ .

El proceso de Monte Carlo se realiza de la siguiente manera:

Decimos que  $PSL(L_A)$  es la distribución de los  $PSL_i$  para una formación  $L$  del equipo  $A$ . De la formación  $L_A$  podemos construir la matriz de variables aleatorias  $\mathbf{R}$ . Luego:

<b>Input:</b> Número de simulaciones $N$
<b>Input:</b> Matriz de variables aleatorias $\mathbf{R}$
<b>Output:</b> Distribución del PSL del equipo $A$
<b>1</b> Inicializar $N$ simulaciones;
<b>2</b> $PSL_i \leftarrow 0$ para $i = 1, 2, \dots, N$ ;
<b>3</b> <b>for</b> $i = 1$ <b>to</b> $N$ <b>do</b>
<b>4</b> $R \leftarrow$ Muestrear de la matriz $\mathbf{R}$ distribuciones a priori de los $r(U, V)$ ;
<b>5</b> $Q \leftarrow$ Normalizar las filas de $R$ ;
<b>6</b> $PSL_i \leftarrow PSL(Q)$ ;
<b>7</b> <b>end</b>
<b>8</b> Estimar la distribución del PSL del equipo $A$ a partir de las $N$ observaciones obtenidas de las simulaciones;

**Algorithm 1:** Simulación del PSL del equipo  $A$

A partir de esta distribución del PSL, se puede realizar comparaciones entre diferentes formaciones de 11 jugadores.

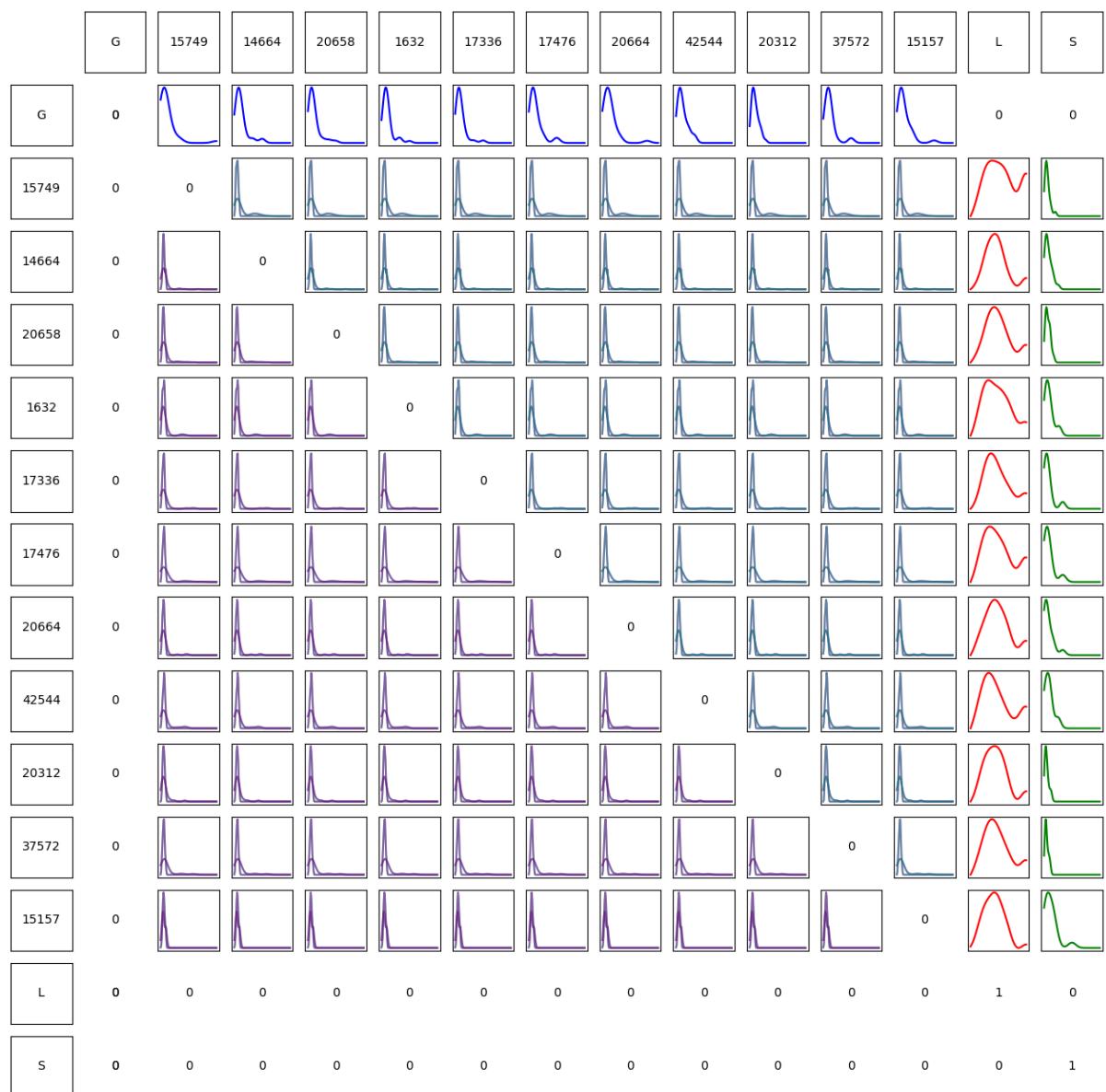


Figure 15: Matriz de Variables Aleatorias  $\mathbf{R}$

El siguiente gráfico muestra la distribución del PSL de un equipo de ejemplo obtenida a partir de 1000 simulaciones del proceso de Monte Carlo para la formación mas utilizada en la temporada 2012/13 de la EPL del equipo Manchester City.

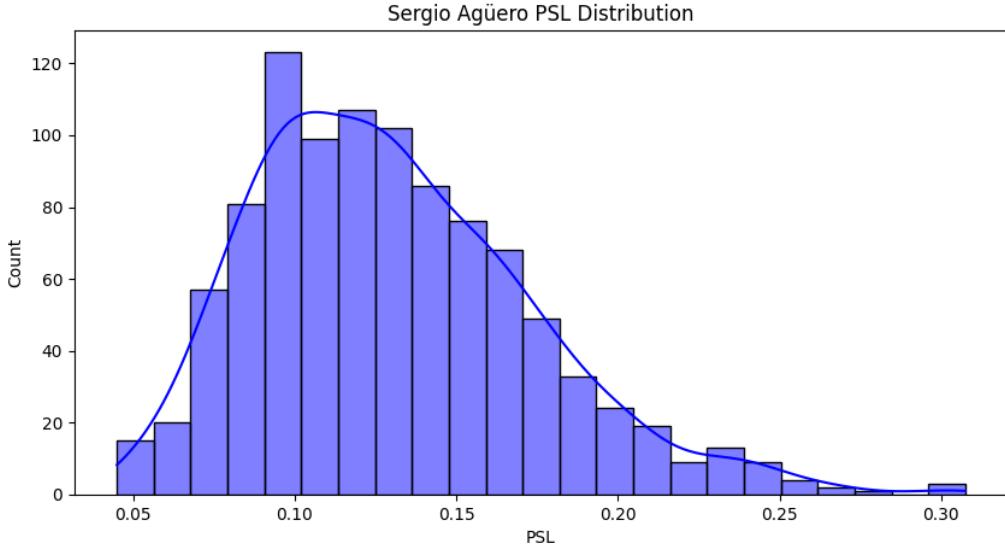


Figure 16: Distribución del PSL del equipo Manchester City

### 0.7.3 Comparar el impacto sobre el PSL de dos jugadores en una formación

Para comparar el PSL de dos jugadores en una formación, se propone un análisis de sensibilidad que consiste en evaluar el impacto en la distribución del PSL al reemplazar a un jugador por otro en la formación. El proceso para ello es el siguiente:

- Input:** Formación  $L_A = \{J_1, J_2, \dots, J_{11}\}$ , Donde algún  $J_i$  es el jugador a "original"

**Input:** Jugador  $J'$  a comparar con  $J_i$

**Input:** Formación  $L'_A = \{J_1, J_2, \dots, J_{11}\}$  **Output:** Distribuciones del PSL de  $L_A$  con  $J_i$  y de  $L'_A$

  - 1 Inicializar  $N$  simulaciones;
  - 2 Estimar la distribución del PSL del equipo  $A$  al reemplazar a  $J_1$  por  $J_2$  a partir de las  $N$  observaciones obtenidas de las simulaciones;

**Algorithm 2:** Comparación de Distribuciones de PSL

## 0.8 Player2Vec: Embeddings de Jugadores

Para poder representar a cada jugador de forma vectorial, se desarrolló el modelo de Player2Vec que permite obtener un embedding de cada jugador en un espacio de  $n$  dimensiones. Un embedding es una representación numérica de objetos en un espacio de  $n$  dimensiones, donde propiedades o relaciones similares se preservan. En el contexto de jugadores, un embedding transforma las características de cada jugador en un vector de números, de tal manera que jugadores con comportamientos o atributos similares estén más cerca en este espacio vectorial. Esto facilita que modelos como redes neuronales aprendan patrones complejos a partir de estas representaciones compactas. ##### Definición

Player2Vec es una adaptación de Node2Vec para representar jugadores de fútbol en un espacio vectorial. En este caso, los nodos del grafo representan jugadores, y las aristas entre ellos reflejan la interacción entre los jugadores en partidos de fútbol. A partir de los datos de eventos de partidos (pases, disparos, goles, etc.), se construye un grafo donde los nodos son jugadores y las aristas representan la frecuencia de interacción entre ellos.

### 0.8.1 Modelado de la EPL 2012/13 como Grafo

A partir de una formación de 11 (Lineup), para un equipo (Team), en un partido (Match), se construye el grafo de la red de jugadores. Llámese a estos  $G_{L,T,M}$  Grafo de Lineup.

Sean:

- $l \in L = \{0, 3\}$  las formaciones posibles (en la temporada 12/13 se permitían hasta 3 cambios de jugadores)
- $t \in T = \{\text{Local, Visitante}\}$  los equipos que jugaron el partido.
- $m \in M = \{1, 2, \dots, 380\}$  los partidos de la temporada 12/13 de la EPL

$$G_{L,T,M} = (V^{L,T,M}, E^{L,T,M})$$

$L$  = Número de Lineup del equipo en el partido

$T$  = Número de Equipo

$M$  = Número de Partido

$$V^{L,T,M} = \{\text{Gain}^{L,T,M}, J_1^{L,T,M}, J_2^{L,T,M}, \dots, J_{11}^{L,T,M}, \text{Loss}^{L,T,M}, \text{Shot}^{L,T,M}\}$$

$$E^{L,T,M} = \{(J_i^{L,T,M}, J_j^{L,T,M}, r(J_i^{L,T,M}, J_j^{L,T,M})) \mid i, j \in [1, 11]\}$$

$$\cup \{(\text{Gain}^{L,T,M}, J_i^{L,T,M}, r(\text{Gain}^{L,T,M}, J_i^{L,T,M})) \mid i \in [1, 11]\}$$

$$\cup \{(J_i^{L,T,M}, \text{Shot}^{L,T,M}, r(J_i^{L,T,M}, \text{Shot}^{L,T,M})) \mid i \in [1, 11]\}$$

$$\cup \{(J_i^{L,T,M}, \text{Loss}^{L,T,M}, r(J_i^{L,T,M}, \text{Loss}^{L,T,M})) \mid i \in [1, 11]\}$$

Donde cada  $J_i^{L,T,M} \mid i \in [1, 11]$  es un nodo que representa a un jugador en el lineup  $L$  del equipo  $T$  en el partido  $M$ .  $\text{Gain}^{L,T,M}$  es el nodo que representa la ganancia del balón,  $\text{Loss}^{L,T,M}$  la pérdida del balón y  $\text{Shot}^{L,T,M}$  el disparo al arco en el lineup  $L$  del equipo  $T$  en el partido  $M$ .

En la figura se visualiza un ejemplo de un grafo de lineup  $G^{L,T,M}$  genérico con los ejes  $r(J_1^{L,T,M}, U)$  resaltados.

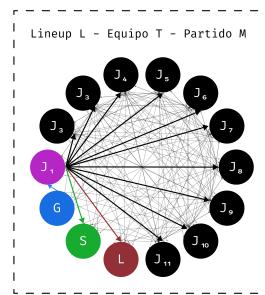


Figure 17: Grafo de Lineup

Luego sean: -  $J_i \mid i \in [0, 522]$  los jugadores reales de la temporada 2012/13 de la EPL

Se construye el grafo de la red de jugadores  $G_{\text{EPL-12/13}}$  como la unión de todos los grafos de lineup  $G^{L,T,M}$ .

$$\begin{aligned}
G_{\text{Full}} = (V, E) &= \bigcup_{L,T,M} G^{L,T,M} \\
V &= \{J_1, J_2, \dots, J_{522}, \text{Gain}, \text{Loss}, \text{Shot}\} \\
&\cup \bigcup_{L,T,M} \{J_1^{L,T,M}, J_2^{L,T,M}, \dots, J_{11}^{L,T,M}, G^{L,T,M}, L^{L,T,M}, S^{L,T,M}\} \\
E &= \bigcup_{L,T,M} E^{L,T,M} \\
&\cup \{(J_i, J_j^{L,T,M}, r(J_i, J_j^{L,T,M})) \mid i \in [0, 522], j \in [1, 11], L, T, M\} \\
&\cup \{(\text{Gain}, \text{Gain}^{L,T,M}, 1) \mid L, T, M\} \\
&\cup \{(\text{Loss}^{L,T,M}, \text{Loss}, 1) \mid L, T, M\} \\
&\cup \{(\text{Shot}^{L,T,M}, \text{Shot}, 1) \mid L, T, M\}
\end{aligned}$$

El ratio de transición  $r(J_i, J_i^{L,T,M})$  es el tiempo jugado por el Jugador  $J_i$  en el lineup  $L$  del equipo  $T$  en el partido  $M$  sobre el tiempo total jugado por el Jugador  $J_i$

$$r(J_i, J_i^{L,T,M}) = \frac{\text{Time Played}_{J_i^{L,T,M}}}{\text{Time Played}_{J_i}}$$

La siguiente figura es una visualización de una instancia de un Equipo en un Partido con sus lineups. En este caso el equipo hizo dos cambios en el partido ( $J_4$  por  $J_{12}$  y  $J_2$  por  $J_{13}$ ). Se puede observar como los jugadores reales  $J_4$  y  $J_{12}$  se encuentran representados por el mismo nodo  $J_4^{L,T,M}$  y lo mismo para  $J_2$  y  $J_{13}$  con  $J_2^{L,T,M}$  para sus respectivos lineups. El resto de los nodos de jugadores reales mantienen su identidad en los grafos de lineups.

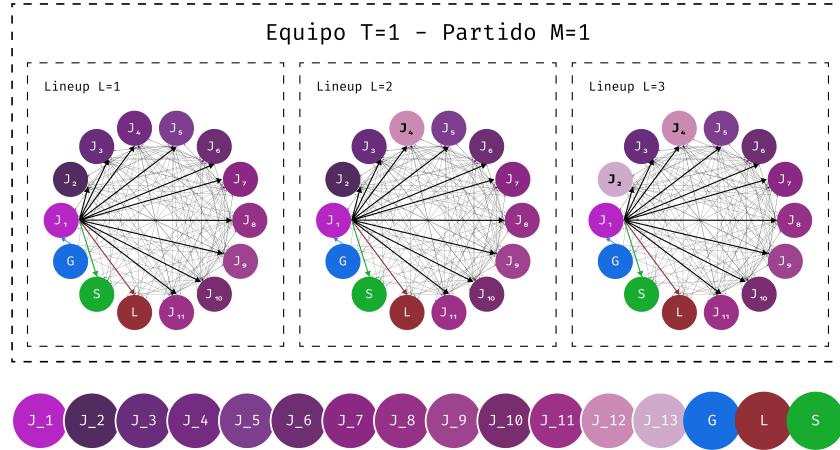


Figure 18: Grafo de Jugadores

El grafo resultante de la composición de todos los grafos de lineup  $G_{\text{Full}}$  se puede comprender mejor en la siguiente visualización:

Donde al igual que en la figura anterior, los nodos de jugadores reales se encuentran representados por los nodos de los lineups en los que participaron.

### 0.8.2 Implementación

A partir de calcular las matrices de ratios  $R^{L,T,M}$  para cada lineup  $L$  del equipo  $T$  en el partido  $M$  generamos el grafo dirigido  $G^{L,T,M}$  haciendo uso de la librería NetworkX en Python para luego componerlos

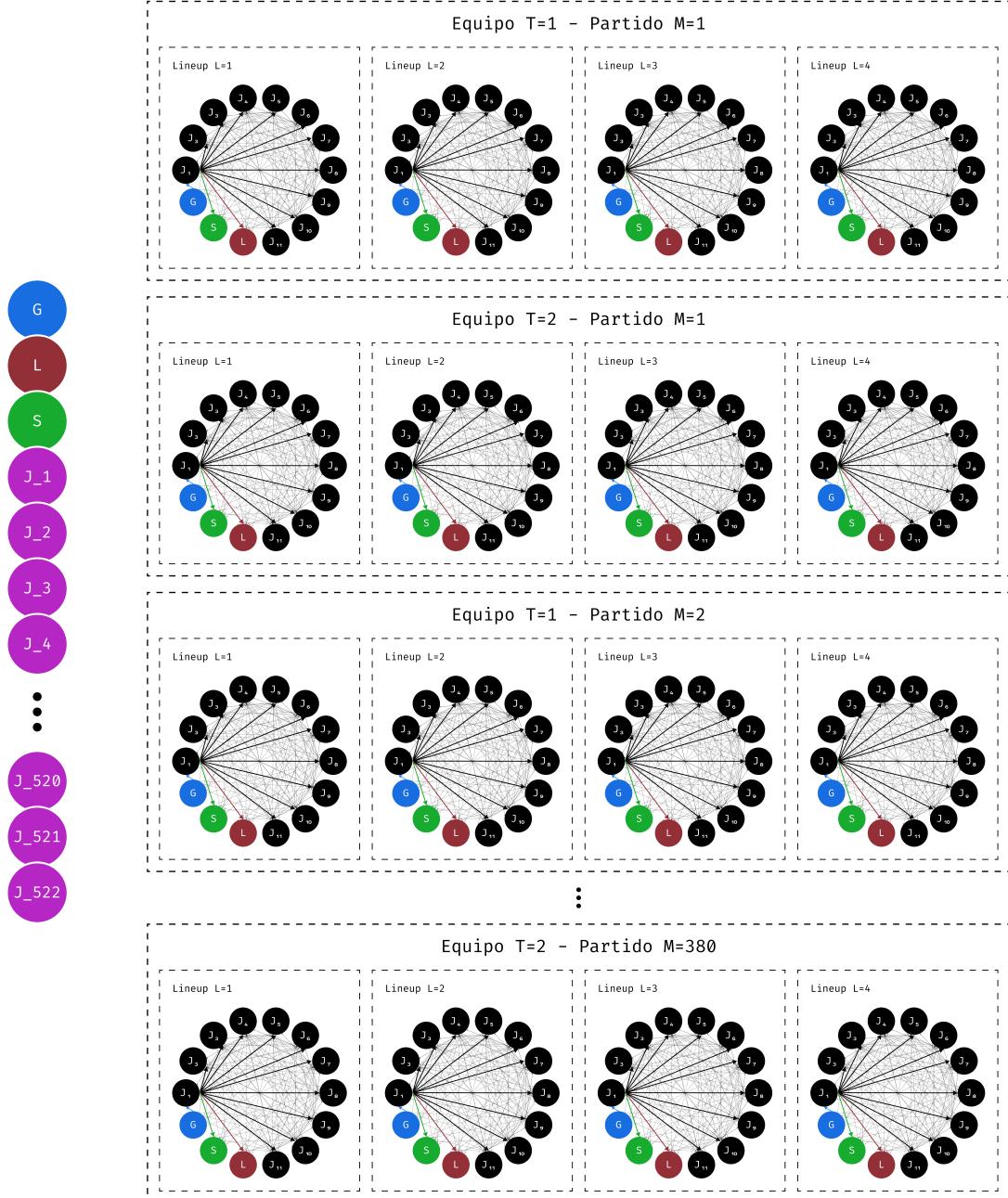


Figure 19: Grafo de Jugadores Completo

en  $G_{\text{Full}}$ , el grafo resultante contiene 37521 nodos y 47338 aristas.

Para obtener los embeddings de los jugadores, se utilizó la librería `node2vec` en Python, que implementa el algoritmo homónimo. Se configuró el modelo con una longitud de caminata de 16 nodos, 200 caminatas y un tamaño de ventana de 12 nodos. Se entrenaron 2 modelos de embeddings, uno con 64 dimensiones para utilizar en modelos de Deep Learning y otro con 3 dimensiones.

Para cada uno de los 37521 nodos se obtuvo un embedding, de los cuales nos quedamos solo con los 522 embeddings de los jugadores reales, estos finalmente son la representación vectorial de cada jugador en el espacio de embeddings.

Este modelo hace uso de Node2Vec, que es en sí una adaptación de Word2Vec, una técnica de NLP que permite representar palabras en un espacio vectorial (Grover & Leskovec, 2016; Mikolov et al., 2013).

Node2Vec es un algoritmo que aprende representaciones vectoriales (embeddings) para nodos en un grafo, preservando tanto las relaciones locales como las globales entre ellos. Utiliza técnicas de random walks para capturar el contexto de cada nodo, balanceando entre explorar nodos cercanos y lejanos. Estos embeddings son útiles para tareas de machine learning sobre grafos, ya que capturan de forma eficiente las interacciones entre nodos en el grafo.

En el caso de Player2Vec, los  $k$  random walks resultantes son una secuencia de jugadores y/o estados de juego en un partido de fútbol (Ganancia, Pérdida, Disparo). A modo ilustrativo los siguientes son posibles random walks obtenidos del grafo de la EPL 2012/13:

$$\begin{aligned} \text{Random Walk 1: } & \text{Gain} \rightarrow \text{Gain}^{L,T,M} \rightarrow J_1^{L,T,M} \rightarrow J_7^{L,T,M} \rightarrow \dots \rightarrow \text{Shot}^{L,T,M} \rightarrow \text{Shot} \\ \text{Random Walk 2: } & J_{93} \rightarrow J_{93}^{L,T,M} \rightarrow J_{15}^{L,T,M} \rightarrow J_{21}^{L,T,M} \rightarrow \text{Loss}^{L,T,M} \rightarrow \text{Loss} \\ & \vdots \\ \text{Random Walk } k: & J_{12} \rightarrow J_{12}^{L,T,M} \rightarrow J_{13}^{L,T,M} \rightarrow J_{33}^{L,T,M} \rightarrow \text{Shot}^{L,T,M} \rightarrow \text{Shot} \end{aligned}$$

La cantidad de random walks  $k$  así como los otros hiperparametros del modelo de Node2Vec fueron seleccionados de forma empírica observando el resultado de los embeddings obtenidos.

### 0.8.3 Visualización y Exploración de los Embeddings

Para comenzar a explorar el espacio vectorial generado por Player2Vec, se ajustó un modelo inicialmente a partir siguientes hiperparametros:

- Dimensión de embeddings: 3
- Longitud de caminata: 16 nodos
- Número de caminatas: 200
- Tamaño de ventana: 12 nodos

Se entrenó el modelo y se obtuvieron los embeddings de los 522 jugadores de la temporada 2012/13 de la EPL. La siguiente visualización muestra los embeddings de los jugadores en un espacio de 3 dimensiones, el color corresponde al equipo en el que juega el jugador.

Para poder visualizar de forma más clara los embeddings de los jugadores, se realizó un PCA para reducir la dimensionalidad de los embeddings a 2 dimensiones. La siguiente visualización muestra los embeddings de los jugadores en un espacio de 2 dimensiones, el color corresponde al equipo en el que juega el jugador.

En la figura de los componentes principales se observa como los jugadores de un mismo equipo se encuentran cercanos en el espacio vectorial, lo que indica que los embeddings resultantes de este modelo capturan las relaciones entre los jugadores de un mismo equipo.

Ademas se observa como en este espacio las direcciones en las que se representan a los equipos divergen de forma clara, lo que indica que los embeddings capturan únicamente las diferencias entre los equipos y no las similitudes. Buscan cierta ortogonalidad entre los equipos que no logra existir en este espacio de 3 dimensiones.

Para explotar aún mas las relaciones a aprender por el modelo, se ajustó un segundo modelo con las siguientes características:

Embeddings de 3 dimensiones de los jugadores - Player2Vec

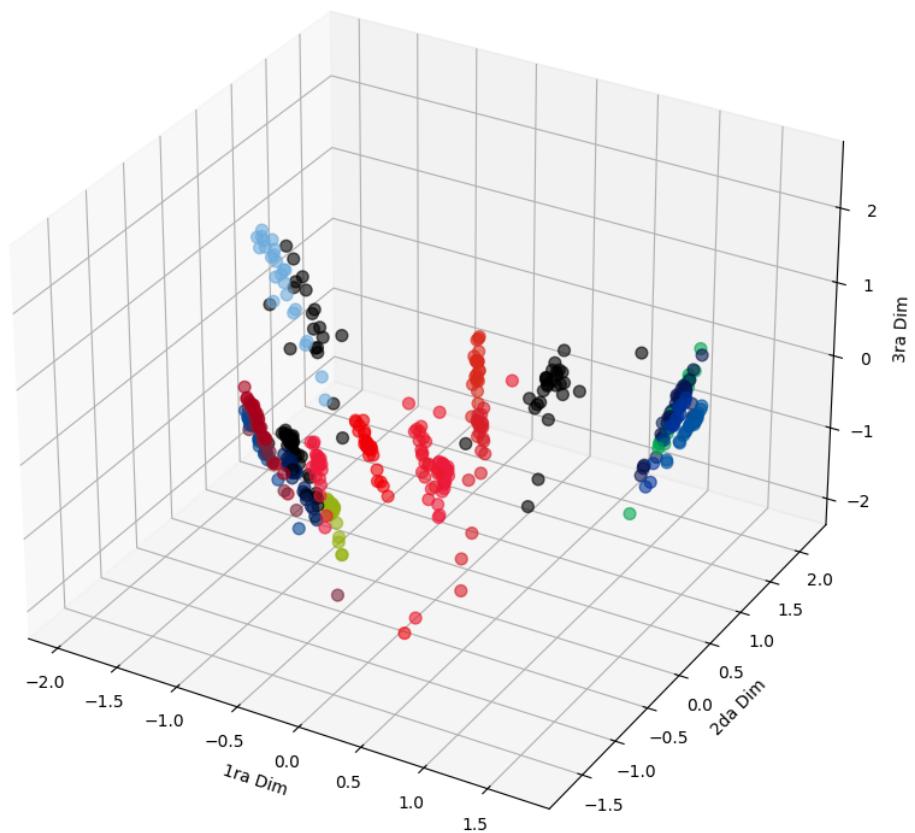


Figure 20: Embeddings de Jugadores en 3D

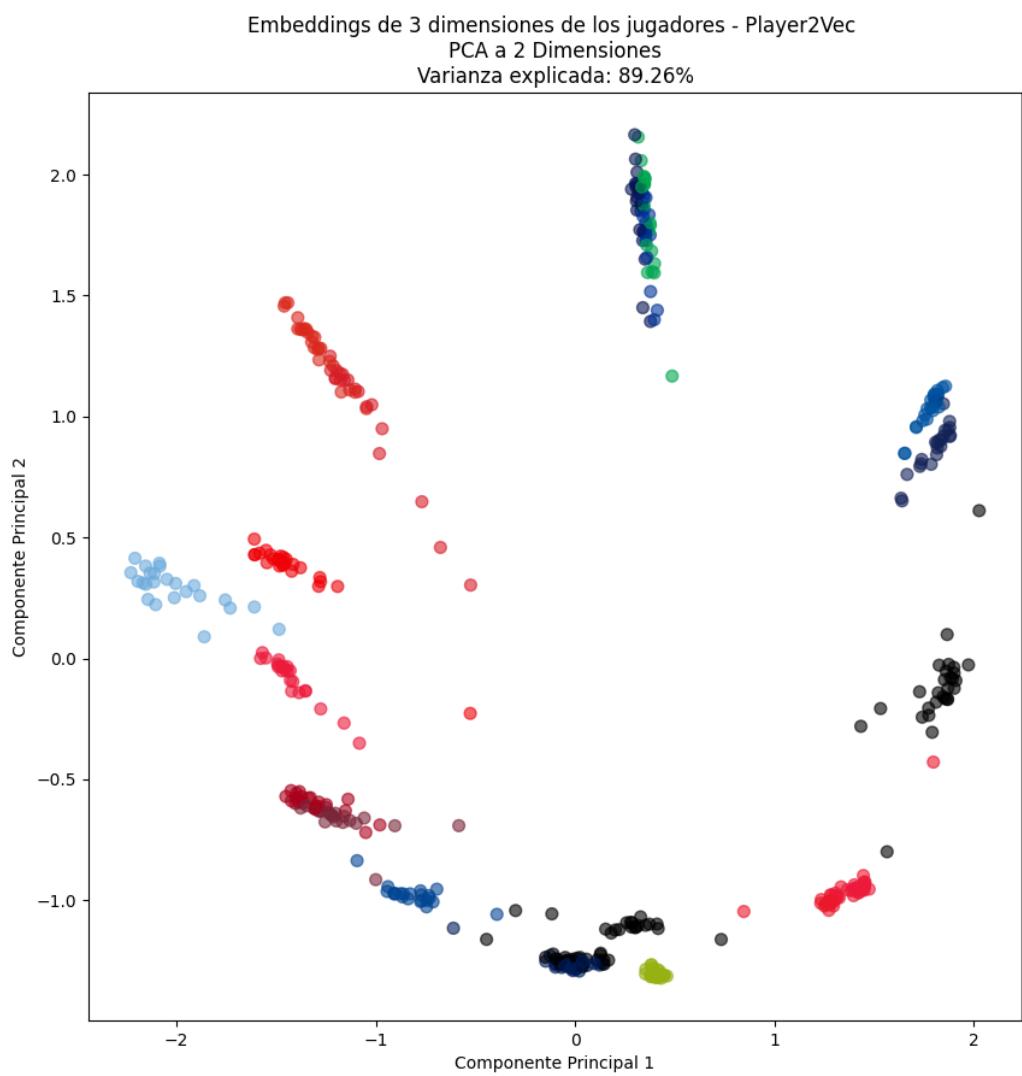


Figure 21: Embeddings de Jugadores en 3D - PCA a 2D

- Dimensión de embeddings: 64
- Longitud de caminata: 40 nodos
- Número de caminatas: 500
- Tamaño de ventana: 30 nodos

Luego para explorar los embeddings resultantes se realizó nuevamente un análisis de componentes principales para reducir la dimensionalidad de los embeddings a 2 dimensiones.

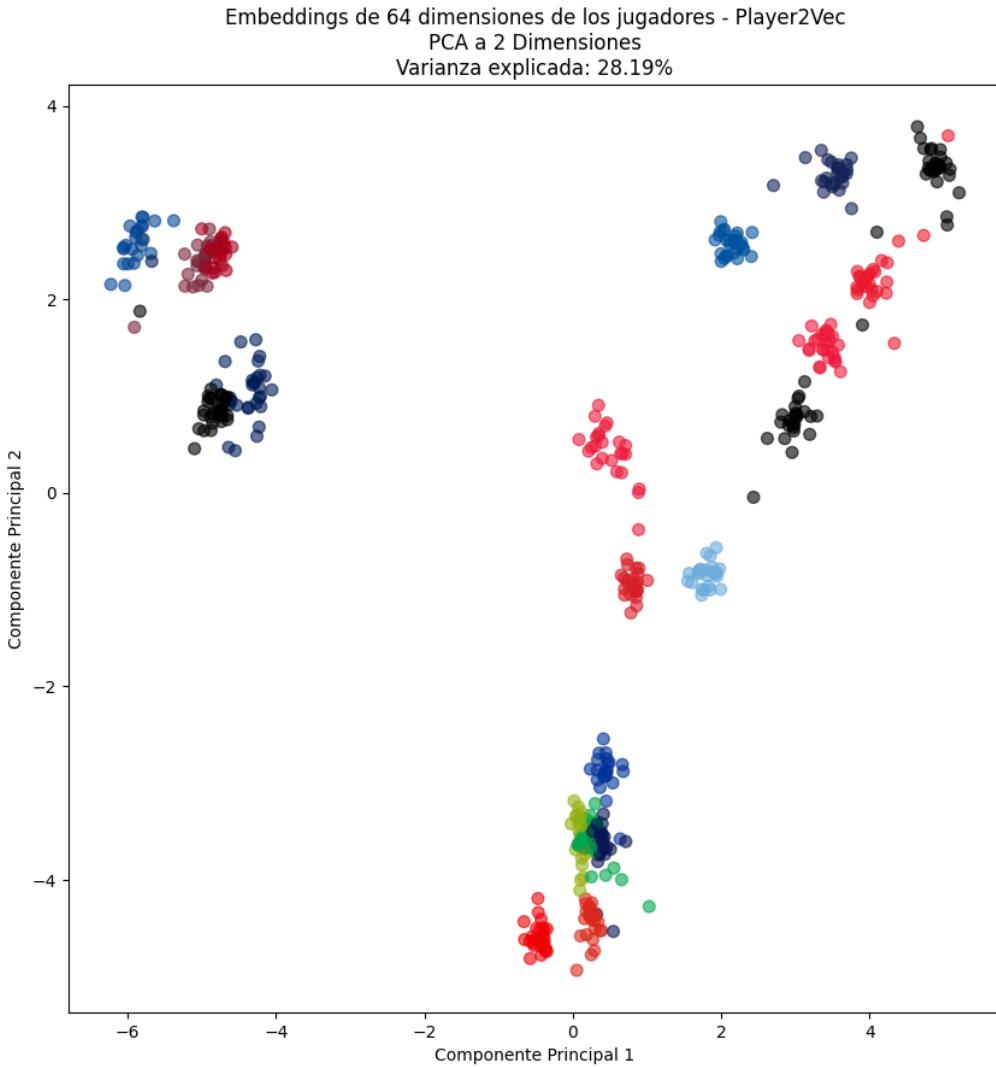


Figure 22: Embeddings de Jugadores en 64D - PCA a 2D

En esta figura resultante se puede observar como la direccionalidad de los equipos desaparece pero se mantienen las relaciones entre los jugadores de un mismo equipo.

#### 0.8.4 Potencial de Player2Vec

Con el modelo planteado de Grafos de Lineups por Equipos y Partidos se puede representar no solo una temporada de una liga, como es nuestro caso, sino que se puede extender a múltiples temporadas y ligas. Esto permitiría poder comparar jugadores de distintas ligas y temporadas, y poder evaluar el rendimiento de un jugador en distintos contextos.

Otra cuestión considerada para expandir es además de tener un nodo general por jugador conectado a sus

instancias en cada lineup, se podría tener un nodo que represente a un jugador en un equipo, de forma tal que el jugador real esté conectado a su nodo “Jugador en Equipo” y este nodo a su vez conectado a “Jugador en Lineup de Partido de Equipo”. Esto permitiría poder evaluar el rendimiento de un jugador en un equipo en particular y cómo este se comporta en distintos contextos.

En el paper de *Soccer Networks* donde se plantea el PSL definen una serie de coeficientes  $h$ ,  $a$ ,  $\omega$ , como la performance de un equipo al jugar de local, al jugar de visitante, y la performance ponderada de todos los otros equipos al jugar de visitante respectivamente. Se podría escalar por ejemplo los ratios de transición entre jugadores y el estado de disparo al arco en función de estos coeficientes para obtener una mejor representación de la performance de un jugador en un partido en particular.

## 0.9 Hipótesis

En los casos en que corresponda, la hipótesis es una propuesta que se someterá a prueba a lo largo de la investigación, basada en el planteamiento del problema.

## 0.10 Marco teórico

- **Proyectos de Investigación:** Aquí se exponen las teorías, antecedentes y conceptos clave relacionados con el tema de estudio. Sirve para fundamentar el trabajo con bases teóricas y estudios previos. Se deben incluir metodologías de desarrollo de software (ágil, Scrum, etc.), paradigmas de programación, tecnologías y frameworks utilizados (por ejemplo, React, Node.js, bases de datos SQL o NoSQL)
- **Proyectos de Desarrollo:** En esta sección, se debe exponer la base conceptual y técnica que sustenta el proyecto. Esto incluye una revisión de metodologías de desarrollo de software (ágil, Scrum, etc.), paradigmas de programación, tecnologías y frameworks utilizados (por ejemplo, React, Node.js, bases de datos SQL o NoSQL). También se pueden mencionar investigaciones previas o estudios de casos relevantes, siempre enfocándose en cómo estas teorías y tecnologías se relacionan con el proyecto.

## 0.11 Marco metodológico

- **Proyectos de Investigación:** Describe el enfoque metodológico utilizado en la investigación, es decir, cómo se va a llevar a cabo el estudio. Incluye el diseño de la investigación, las técnicas de limpieza de datos (si hubo) y el análisis que se hará. Se debe incluir un flujo o diagrama de la arquitectura de la solución planteada.
- **Proyectos de Desarrollo:** Describe el proceso de desarrollo seguido durante el proyecto. Especifica cómo se organizaron los sprints, las iteraciones, o las fases del proyecto. Incluye también las herramientas empleadas para la gestión del proyecto (Jira, Trello, GitHub) y el enfoque en las pruebas de software para asegurar la calidad del producto final. Se debe incluir un flujo o diagrama de la arquitectura de la solución planteada.

## 0.12 Resultados

- **Proyectos de Investigación:** Se presentan los datos o hallazgos obtenidos en la investigación de forma clara y organizada. En este apartado no se deben interpretar los resultados, sólo exponerlos.
- **Proyectos de Desarrollo:** Aquí se detallan los entregables del proyecto de software. Esto puede incluir versiones funcionales del software, demostraciones de características clave, documentación técnica, y análisis de los tiempos de desarrollo o eficiencia lograda. También se pueden incluir métricas como la satisfacción del cliente, el rendimiento del sistema, la escalabilidad o la compatibilidad multiplataforma, según la naturaleza del software desarrollado.

## 0.13 Discusión

- **Proyectos de Investigación:** Aquí se interpretan los resultados obtenidos, comparándolos con la literatura revisada en el marco teórico. Se analizan los hallazgos y se discute su relevancia.

- **Proyectos de Desarrollo:** En esta sección, se realiza un análisis crítico de los resultados obtenidos. Compara las metas iniciales del proyecto con los entregables finales y evalúa si las expectativas fueron cumplidas. Reflexiona sobre los desafíos técnicos encontrados, como problemas de compatibilidad, rendimiento, o integración de APIs, y cómo fueron solucionados. Además, discute el impacto potencial del software en la empresa privada para la cual fue desarrollado, evaluando su viabilidad, utilidad, y posibles mejoras para futuras versiones.

## 0.14 Conclusiones & Recomendaciones {#conclusiones-&-recomendaciones}

Se sintetizan los puntos más importantes del estudio, haciendo énfasis en si se cumplieron los objetivos y qué se aprendió a partir de los resultados obtenidos.

También, se sugieren posibles líneas de investigación futuras o iteraciones del proyecto, aplicaciones prácticas de los resultados o recomendaciones para la implementación de los hallazgos en la vida real.

## 0.15 Referencias bibliográficas

- Grover, A., & Leskovec, J. (2016). *node2vec: Scalable feature learning for networks*. arXiv.org. <https://arxiv.org/abs/1607.00653>
- Huang, E., Segarra, S., Gallino, S., & Ribeiro, A. (n.d.). *How to find the right player for your soccer team?*
- Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). *Efficient estimation of word representations in vector space*. arXiv.org. <https://arxiv.org/abs/1301.3781>
- Rahimian, P., Van Haaren, J., & Toka, L. (2023). Towards maximizing expected possession outcome in soccer. *International Journal of Sports Science & Coaching*, 174795412311544. <https://doi.org/10.1177/17479541231154494>

## 0.16 Apéndices: Tablas, figuras, anexos {#apéndices:-tablas,-figuras,-anexos}

Se incluyen materiales adicionales como gráficos, tablas, cuestionarios, o documentos que sean relevantes pero no forman parte del cuerpo principal de la tesis.

## Índice de Figuras

1	Modelo de Red de Jugadores . . . . .	5
2	Resultados Modelo de Regresión Lineal . . . . .	7
3	Resultados Modelo de XGBoost . . . . .	8
4	Distribución de todos los $r(J, S)$ . . . . .	9
5	Distribución de los $r(J, S)$ de Sergio Agüero y Robin van Persie . . . . .	10
6	Distribución de los $r(J, S)$ de los 10 jugadores con mayor cantidad de disparos . . . . .	10
7	Distribución de los $r(J, S)$ de los 10 jugadores con mayor sesgo . . . . .	10
8	Top 20 Delanteros con distribución Beta más sesgada a la derecha - EPL 12/13 . . . . .	11
9	Distribución de los $r(J, S)$ de los 10 jugadores con mayor suma . . . . .	11
10	Distribución de los $r(J, S)$ de jugadores similares a Sergio Agüero . . . . .	12
11	Distribución de los $r(J, S)$ de jugadores similares a Sergio Agüero Superpuestos . . . . .	12
12	Distribución de los $r(J, S)$ de jugadores similares a Sergio Agüero de la misma posición . . . . .	12
13	Distribución de los parámetros $\alpha$ y $\beta$ de los $r(J, S)$ de los jugadores . . . . .	13
14	Distribución de los $r(J, S)$ de jugadores en clusters . . . . .	13
15	Matriz de Variables Aleatorias <b>R</b> . . . . .	15
16	Distribución del PSL del equipo Manchester City . . . . .	16
17	Grafo de Lineup . . . . .	17
18	Grafo de Jugadores . . . . .	18
19	Grafo de Jugadores Completo . . . . .	19
20	Embeddings de Jugadores en 3D . . . . .	21
21	Embeddings de Jugadores en 3D - PCA a 2D . . . . .	22
22	Embeddings de Jugadores en 64D - PCA a 2D . . . . .	23

## Índice de Tablas