Puntajes del videojuego FIFA como modelo matemático para la estimación del rendimiento en el fútbol profesional

David González Escobar

Trabajo de Grado

Asesor:
Andrés Ramírez Hassan

Universidad EAFIT
Mayo 2021

Abstract

La analítica de datos se ha convertido en una herramienta que permite apoyar la toma de decisiones en diversos deportes. Entre estos, el fútbol se encuentra rezagado. En este artículo, se exploran los puntajes del videojuego FIFA como una posible herramienta para cerrar esta brecha. Se utilizan modelos de regresión Poisson Bivariados – incluyendo y sin incluir variables derivadas de FIFA - para estimar las probabilidades de posibles resultados de la liga colombiana 2017. Se evalúa el porcentaje de acierto frente a los resultados reales y los retornos a un portafolio de apuestas siguiendo las predicciones de los modelos. Se concluye que el modelo que incluye variables del videojuego FIFA tiene un mayor predictivo y genera mejores retornos.

Introducción

Es famosa la anécdota que Andrea Pirlo – legendario mediocampista de la selección italiana de fútbol – en donde tomó una decisión sumamente importante en su preparación para jugar la final del Campeonato Mundial de la FIFA en 2006: pasar la mañana jugando fútbol en Playstation. Así lo hizo. En la noche quedó campeón de la competición más importante que existe de este deporte.

Los videojuegos deportivos son el entretenimiento de millones de personas alrededor del mundo. El videojuego FIFA de EA Sports – que desde su lanzamiento en el año 1993 saca una edición anual en diversas consolas – ha logrado posicionarse como el juego de fútbol de referencia en el mercado.

Para esto, además de poseer licencias exclusivas de equipos y jugadores de la FIFA, han ido mejorando año a año la calidad de la experiencia que le dan a sus usuarios, buscando que lo que ocurre en la pantalla cada vez se asemeje más a lo que ocurre en el mundo real. De esta forma, los videojuegos FIFA han logrado – paralelamente a su éxito comercial - algo sumamente valioso: ser un modelo matemático para lo que ocurre en el campo de juego.

Los videojuegos FIFA asignan puntajes numéricos a cada jugador en diversos aspectos del juego como su velocidad máxima y aceleración, su precisión a la hora de realizar un pase o la fortaleza

de sus remates al arco. Con esto, buscan que sus habilidades en la vida real se traduzcan a un desempeño equivalente en el mundo virtual.

Estos puntajes numéricos son definidos para cada edición del juego por más de 9.000 personas, y llegan a ser diferenciadores en la forma en la cual un jugador se desempeña en el videojuego que hasta jugadores profesionales que crecieron jugando FIFA llegan a tenerlos de referencia y a tratar de moldear su juego a acorde a los puntajes que les gustaría tener.

Ahora, además de lograr describir en un mundo simulado el comportamiento de jugadores en el mundo real, ¿serán estos puntajes numéricos capaces de tener algún poder predictivo respecto al desempeño de equipos profesionales de fútbol en el mundo real?

Antecedentes

En los deportes profesionales, al máximo nivel de competitividad y dedicación, los detalles logran marcar la diferencia. Por esto mismo, los equipos y jugadores profesionales están constantemente "innovando", en busca de eso que les pueda dar la ventaja definitiva frente a sus rivales: sea un entrenamiento más exigente para lograr una ventaja física, el desarrollo de habilidades en etapa temprano al capacitar divisiones inferiores o la implementación de una táctica nunca antes vista en un contexto profesional.

En busca de estas "innovaciones", los clubes y jugadores profesionales de distintos deportes han acudido en los últimos años a la revolución del Big Data y las técnicas estadísticas para analizar estos mismos datos, en aras de lograr marcar la diferencia. Este uso de la analítica de datos se podría catalogar dentro de tres categorías: para adquirir talento, para desarrollar talento y para la implementación de estrategias nuevas dentro del campo del juego (Grow, 2020).

Los deportes americanos (béisbol, basquetbol, fútbol americano, etc.) han sido pioneros en la implementación de la analítica para la toma de decisiones de alto alcance para equipos profesionales. El caso más famoso es de los Oakland Athletics y su gerente general Billie Bean en la temporada 2002 de la MLB, que sería popularizado posteriormente en el libro y película

homónima Moneyball. Con un presupuesto limitado y totalmente en contra de la tradicional "sabiduría" intangible del deporte, Beane desarrolló una estrategia para adquirir talento para la plantilla de los Athletics basado en *sabermetrics,* que es el análisis del béisbol a partir de estadísticas. Fijándose únicamente en la métrica de "porcentaje en base (OBS)" en vez de los atributos físicos y de estilo utilizados usualmente por los cazatalentos, Beane logró construir una plantilla competitiva que – contra todo pronóstico – logró una temporada exitosa, de pasó batiendo el inédito record de ganar 19 encuentros de forma consecutiva (Wolfe et al., 2006).

A partir de esto, muchos equipos profesionales en los Estados Unidos crearon departamentos de analítica que sirvieran de apoyo para su toma de decisiones (Freeman, 2016). El caso de Moneyball es una muestra de libro de cómo la analítica de datos puede implementarse a la hora de adquirir talento. Sin embargo, hay ejemplos claros de su uso para los otros objetivos mencionados anteriormente. Los Houston Rockets en la NBA, guiados por el análisis cuantitativo de los "puntos esperados" que generaba cada intento de campo de acuerdo a su posición en la cancha, cambiaron su estrategia de juego radicalmente al enfocarse únicamente en hacer tiros de tres puntos, intentos de tiro en la pintura y tiros libres, obteniendo gran éxito con esta innovación táctica e influyendo en el estilo de juego del resto de la liga (Paine & Herring, 2021).

Los equipos que han incorporado departamentos y técnicas de analítica de datos han visto como resultado una mejora en su desempeño profesional (Grow, 2020). Ahora bien, la implementación de estas nuevas herramientas no ha sido homogénea entre todos los deportes. En particular, el fútbol - a pesar de gozar del estatus de ser el deporte más popular del mundo – se enfrenta a dificultades inherentes a su naturaleza misma a la hora de implementar estas técnicas: es un deporte muy fluido y caótico (Burn-Murdoch, 2018).

Los deportes mencionados anteriormente (béisbol y basquetbol), están estructurados de una forma tal que se pueden segmentar con relativa sencillez a eventos discretos: posesiones, bateos, etc. Esto permite que sea más fácil de estructurar la información que se crea en cada partido para su análisis posterior. Adicionalmente, en deportes como el béisbol – dada la naturaleza de las interacciones del bateador y el lanzador – es más fácil distinguir la relación entre los resultados de una jugada

entre la habilidad del jugador y la aleatoriedad. En el fútbol, donde tantas cosas interactúan y en el que los resultados de éxito usualmente medidos son muy escasos (los goles), esto no ocurre.

Por todo lo anterior, la implementación de herramientas analíticas ha sido lenta en el fútbol (Fernández, 2020). No obstante, se han logrado desarrollar sistemas que provean información más allá de estadísticas simples cómo los goles, porcentaje de pases exitosos o el tiempo de posesión. Compañías como Opta Sports se han enfocado en ofrecer servicios profesionales a clubes al ser capaces de recopilar hasta 2.000 puntos de información distintos de un solo partido (Burn-Murdoch, 2018).

En los últimos años, varios equipos profesionales han incorporado a sus análisis la novedosa métrica de "goles esperados (xG)", una medida de la probabilidad de que – dada la posición de un jugador ofensivo en el campo y la de todos los demás jugadores en ese momento – un jugador logre anotar un gol. De esta forma, los goles esperados buscan dar una medida que desligue la aleatoriedad de lo que ocurre en el campo del juego, siendo un mejor estimador del rendimiento en el largo plazo que medir únicamente los goles, y permitiendo ver qué equipos o jugadores presentan un rendimiento menor o superior al que se hubiera esperado de ellos (Burn-Murdoch, 2018) (Brechot, 2020). Dar puntajes numéricos al rendimiento en un partido basado en la opinión de un comité de expertos, enfoques para dar un valor distinto a cada acción durante una posesión o enfoques probabilísticos para estimar el valor de cada pase, son otras técnicas que se han sugerido utilizar para el análisis estadístico del fútbol (Fernández, 2020). Sin embargo, todo lo mencionado anteriormente comparte las mismas características: es muy costoso – sea en tiempo o recursos – de estimar. O al menos más costoso que los datos que se recopilan en los deportes pioneros en la analítica de datos.

Esto no quiere decir que los equipos de fútbol no puedan beneficiarse de la disponibilidad de datos estructurados para apoyar su toma de decisiones. Liu et al. (2015), utilizando datos del mundial de fútbol de Brasil 2014 e implementando una modelo de regresión logística, encuentran nueve estadísticas tradicionales que se correlacionan positivamente con la probabilidad de ganar un encuentro. Souza et al. (2019), utilizando datos de ocho temporadas de juego en La Liga Española, utilizan un enfoque similar para hallar que los tiros al arco y los tiros de esquina por partido eran

las variables que mejor predecían los puntos por juego que obtendría un equipo. Müller et al. (2017) - en un aspecto más relacionado con la adquisición de talento que los resultados en el campo de juego - utilizan un enfoque basado en datos para estimar el valor de mercado de jugadores profesionales.

Ahora entra la cuestión, ¿serán los puntajes numéricos asignados a cada jugador en el videojuego FIFA una fuente de datos intuitiva que permita robustecer las herramientas que existen actualmente para realizar análisis cuantitativos en el fútbol profesional?

En la literatura ya hay varios ejemplos de usos de esta base datos para realizar análisis estadísticos. Cotta et al. (2016) utilizan datos de FIFA 15 para hacer un análisis descriptivo de dos temas de discusión entre análisis futbolísticos profesionales: las diferencias entre las selecciones Brasil y Alemania 2014 y el estilo de juego distintivo de Barcelona 2013/14. Matano et al. (2018) utilizan varias estadísticas de jugadores de fútbol y estadísticas de FIFA para crear un modelo tipo Plus – Minus, que llaman Adjusted Plus-Minus (APM), tratando de asemejarse a los modelos estadísticos con poder predictivo existentes en ligas americanas como la NBA y MLB. Verstraete et al. (2020) utilizan también datos de FIFA para realizar analítica de datos para el fútbol. Soto-Valero (2017) usa los datos de FIFA para crear un modelo de clasificación multivariado para los jugadores profesionales de fútbol.

Datos

Se cuenta con datos del videojuego FIFA desde las ediciones FIFA 15 hasta la edición FIFA 21, aproximadamente 15.000 observaciones de jugadores para cada edición del juego. Para cada jugador, se cuenta con variables cualitativas (edad, equipo, nacionalidad), variables físicas (estatura y peso) y las variables correspondientes a los puntajes numéricos asignados por el videojuego (velocidad, regates, definición, etc.). Los datos fueron obtenidos de la plataforma Kaggle.

Para las variables a estimar en el modelo, se cuenta con un conjunto de datos con los resultados de los partidos de la liga colombiana para el torneo de Apertura de 2017. Los datos están estructurados

de forma tal que cada observación corresponde a un partido distinto, donde se tiene información de los dos equipos que se enfrentaron y los goles obtenidos por cada uno de estos.

Así, se añade a las observaciones de cada partido variables derivadas de los datos del videojuego FIFA para el año 2017: el puntaje promedio de los jugadores de cada equipo, el puntaje máximo de un jugador de cada equipo y el puntaje máximo del atributo "definición" de un jugador de cada equipo.

Además, para la validación del modelo a estimar, se cuenta con los resultados de la liga colombiana para el torneo de Clausura 2017, estructurados de la misma forma que los resultados del torneo Apertura. Adicionalmente, para cada partido se incluyen las cuotas de apuestas correspondientes, obtenidas del sitio web Odds Portal.

Metodología

Asumiendo que los goles anotados por los equipos en un partido de fútbol siguen una distribución Poisson bivariada, se sigue la metodología propuesta por Karlis & Ntzoufras (2003). Se procede a estimar el modelo:

$$(X_i, Y_i) \sim BP(\lambda_{1i}, \lambda_{2i}, \lambda_{3i}),$$
$$\ln(\lambda_{1i}) = \mu_1 + local + att_{h_i} + \mathrm{def}_{g_i},$$
$$\ln(\lambda_{2i}) = \mu_2 + att_{g_i} + \mathrm{def}_{h_i},$$

Donde X_i y Y_i son los goles anotados por el equipo local (h_i) y el equipo visitante (g_i) respectivamente en el partido *i,* las cuales siguen una distribución Poisson bivariada. λ_1 y λ_2 son los correspondientes goles esperados de cada equipo, y $cov(X, Y) = \lambda_3$ una medida de la dependencia entre las dos variables aleatorias. μ es un parámetro constante y $local$ es un parámetro que captura el efecto de la ventaja de la localía. Finalmente, att_k y def_k encapsulan el desempeño ofensivo y defensivo del equipo *k.*

Una vez estimado este primer modelo de referencia, se incluirán los regresores derivados de los datos del videojuego FIFA para los jugadores de los equipos para los que se realizan las estimaciones, de la forma:

$$(X_i, Y_i) \sim BP(\lambda_{1i}, \lambda_{2i}, \lambda_{3i}),$$
$$\ln(\lambda_{1i}) = \mu_1 + local + \beta_1 \Delta general_i + \beta_2 \Delta max_i + \beta_3 \Delta definicion_i + att_{h_i} + \text{def}_{g_i},$$
$$\ln(\lambda_{2i}) = \mu_2 + att_{g_i} + \text{def}_{h_i},$$

Donde $\Delta general_i$ es la diferencia entre el puntaje promedio en FIFA de los jugadores de cada equipo, Δmax_i es la diferencia entre el puntaje máximo para un jugador de cada equipo y $\Delta definicion_i$ es la diferencia entre el puntaje máximo de "definición" para un jugador de cada equipo, todos para cada partido *i.*

Utilizando el paquete estadístico *bivpois* en R propuesto por Karlis & Ntzoufras (2005), se estiman varias regresiones para cada estructura de modelo: doble-Poisson, Poisson bivariado y modelos Poisson bivariado con inflado diagonal utilizando varias distribuciones.

Posteriormente, se valida la capacidad predictiva de los modelos resultantes. Primero, se calculan las probabilidades que otorgan los modelos a los distintos resultados (victoria local, victoria visitante o empate) de cada partido. Con esto, seleccionamos únicamente los partidos a los cuales el modelo le asigna más de un 0.5 de probabilidad a un resultado particular (alrededor de la mitad de los partidos). Estos partidos luego son comparados con l los resultados de los partidos del torneo 2017 Clausura como validación: se procede a calcular en qué porcentaje de los partidos en los cuales el modelo le otorga más del 0.5 de probabilidad a un resultado particular este acierta.

Se escogen la configuración del modelo inicial y del modelo incluyendo los regresores derivados del videojuego FIFA que acierten el mayor porcentaje de resultados de los datos de validación. Se compara la capacidad predictiva de cada uno de estos modelos viendo cuál de los dos logra tener un mayor porcentaje de acierto.

Finalmente, se simula un ejercicio para un supuesto portafolio de inversión en apuestas según los resultados que arroja el modelo. Se simula un capital de $1.000 que se distribuye en apuestas

individuales de igual valor en cada partido para el cual el modelo asigna más de un 0.5 de probabilidad a un resultado particular, para después ver cuál sería el retorno de estos portafolios de inversión si se apostara en los partidos que el modelo escoge como los más probables. Al final se compara el retorno del supuesto portafolio si se tomaran las decisiones según el modelo que incluye los resultados de FIFA y el modelo que no los incluye.

Resultados

El modelo seleccionado con mayor poder predictivo para el modelo sin utilizar datos del videojuego FIFA fue el modelo Poisson bivariado inflado con ceros. Para el modelo incluyendo regresores de FIFA, fue el modelo Poisson bivariado con inflado diagonal, donde la diagonal sigue una distribución discreta de un solo parámetro. Los parámetros de los modelos seleccionados se presentan en la Tabla 1. Los tipos de modelos seleccionados tienden a ser más precisos en la estimación de posibles empates en los partidos (Karlis & Ntzoufras, 2003).

Se observa que los parámetros correspondientes a los rendimientos ofensivos y defensivos esperados de cada equipo son consistentes entre los modelos. Los mismos equipos comparten un rendimiento superlativo (Atlético Nacional), como un rendimiento malo (Tigres), lo cual correspondió a la realidad de estos equipos en el torneo Apertura de 2017, donde Nacional lideró cómodamente el campeonato y Tigres disputó los puestos más cercanos al descenso. Esto se puede observar visualmente en los Gráficos 1 y 2, donde se invierte el signo de los parámetros defensivos.

	Modelo Sin FIFA:		**Modelo FIFA:**	
	Ataque	*Defensa*	*Ataque*	*Defensa*
América	0.99	0.15	1.05	-0.01
Nacional	1.12	-1.24	1.13	-1.2
Bucaramanga	0.31	-0.12	0.28	-0.29
Cortuluá	0.42	0.58	0.28	0.61
Cali	0.97	0.12	0.88	0.11
Pasto	0.99	0.01	1.11	-0.24
Tolima	0.84	0.44	0.79	0.34
Envigado	0.31	0.24	0.42	0.01
Huila	0.42	0.37	0.38	0.28
Medellín	1.19	0.22	0.84	0.44
Jaguares	0.36	-0.18	0.36	-0.20

Junior	0.73	0.28	0.69	0.36
Equidad	0.34	0.01	0.29	-0.05
Millonarios	1.01	-0.02	1.10	-0.15
Once Caldas	0.23	0.44	0.18	0.38
Patriotas	0.44	0.02	0.34	-0.04
Alianza	0.79	0.23	0.75	-0.26
Santa Fe	0.27	0.01	0.28	0.03
Tigres	-0.22	0.17	-0.3	0.13
$\boldsymbol{\mu_1}$	-0.67		-0.56	
$\boldsymbol{\mu_2}$	-1.29		-1.20	
Intercepto ($\boldsymbol{\lambda_3}$)	-1.72		-1.88	
$\boldsymbol{\lambda_3}$	0.18		0.153	
local	0.62		0.64	
$\boldsymbol{\beta_1}$	-		-0.033	
$\boldsymbol{\beta_2}$	-		0.122	
$\boldsymbol{\beta_3}$	-		-0.056	
AIC	1059.1		1061.42	
BIC	1226.73		1245.02	

Tabla 1. Parámetros modelos seleccionados

En los partidos en los que el modelo le otorga más del 0.5 de probabilidad a un resultado particular, el modelo que no incluye los datos de FIFA acierta a predecir el 49.5% de los resultados del torneo Clausura 2017. En el caso del modelo utilizando los datos derivados del videojuego como regresores, este porcentaje fue del 51,7%. Es decir, las predicciones del mejor modelo que utiliza datos de FIFA son mejores que las del mejor modelo que no los utiliza. En otras palabras: la inclusión de los regresores derivados del videojuego FIFA aumentaron el poder predictivo del modelo frente a los datos de validación.

Adicionalmente, evaluando el retorno para un portafolio de $1.000, nos encontramos con que apostar según las predicciones del modelo sin datos de FIFA nos retornaría un valor de $1409,8 al finalizar la temporada. El modelo con datos de FIFA nos retornaría $1490,1. Es decir, además de tener mayor poder predictivo, el modelo con los datos de FIFA también nos da un mejor retorno que el modelo sin los datos de FIFA, validando su mejor desempeño.

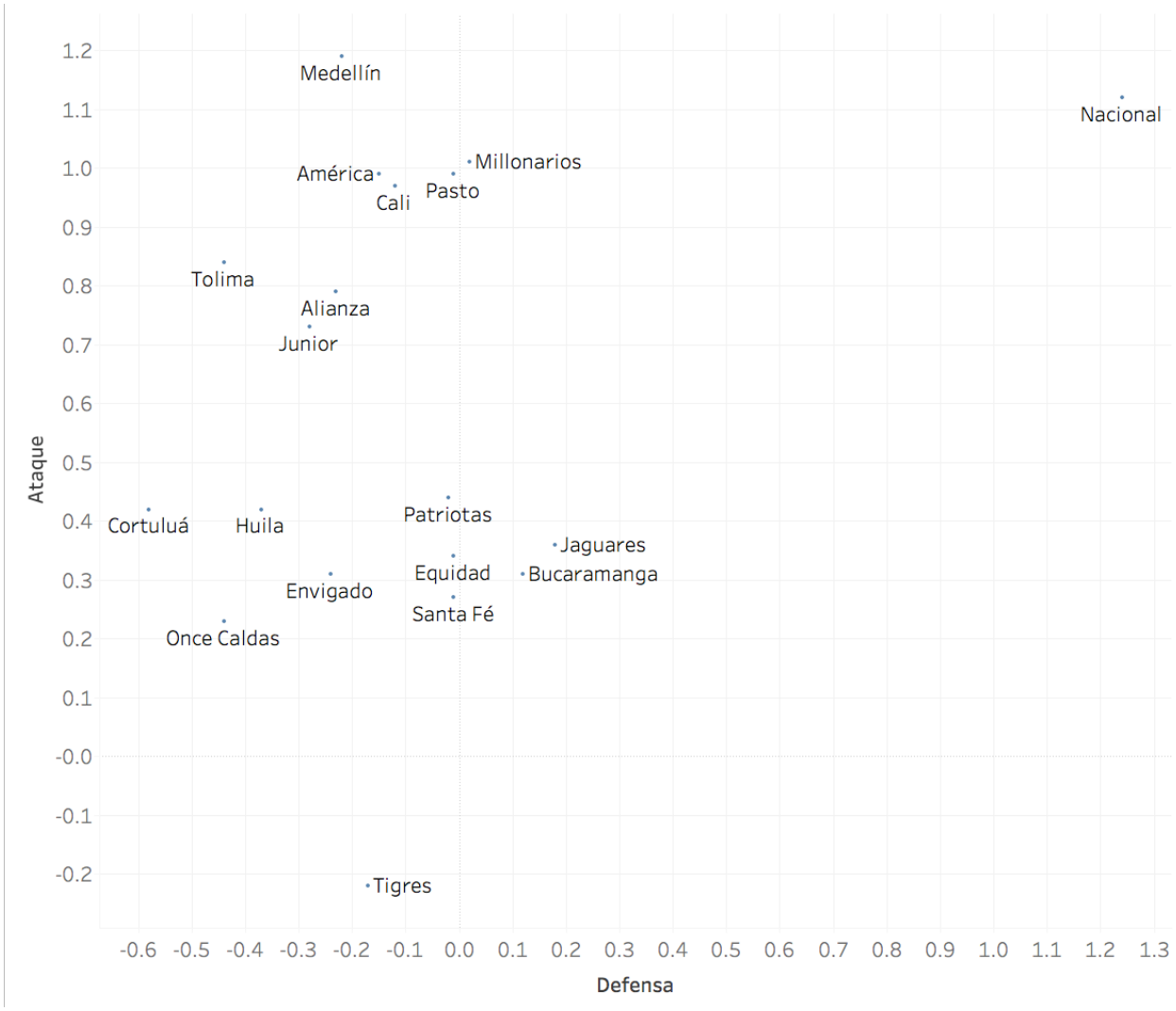

Gráfico 1. Visualización de los parámetros resultantes del modelo sin datos de FIFA (parámetros defensivos se presentan con signos invertidos)

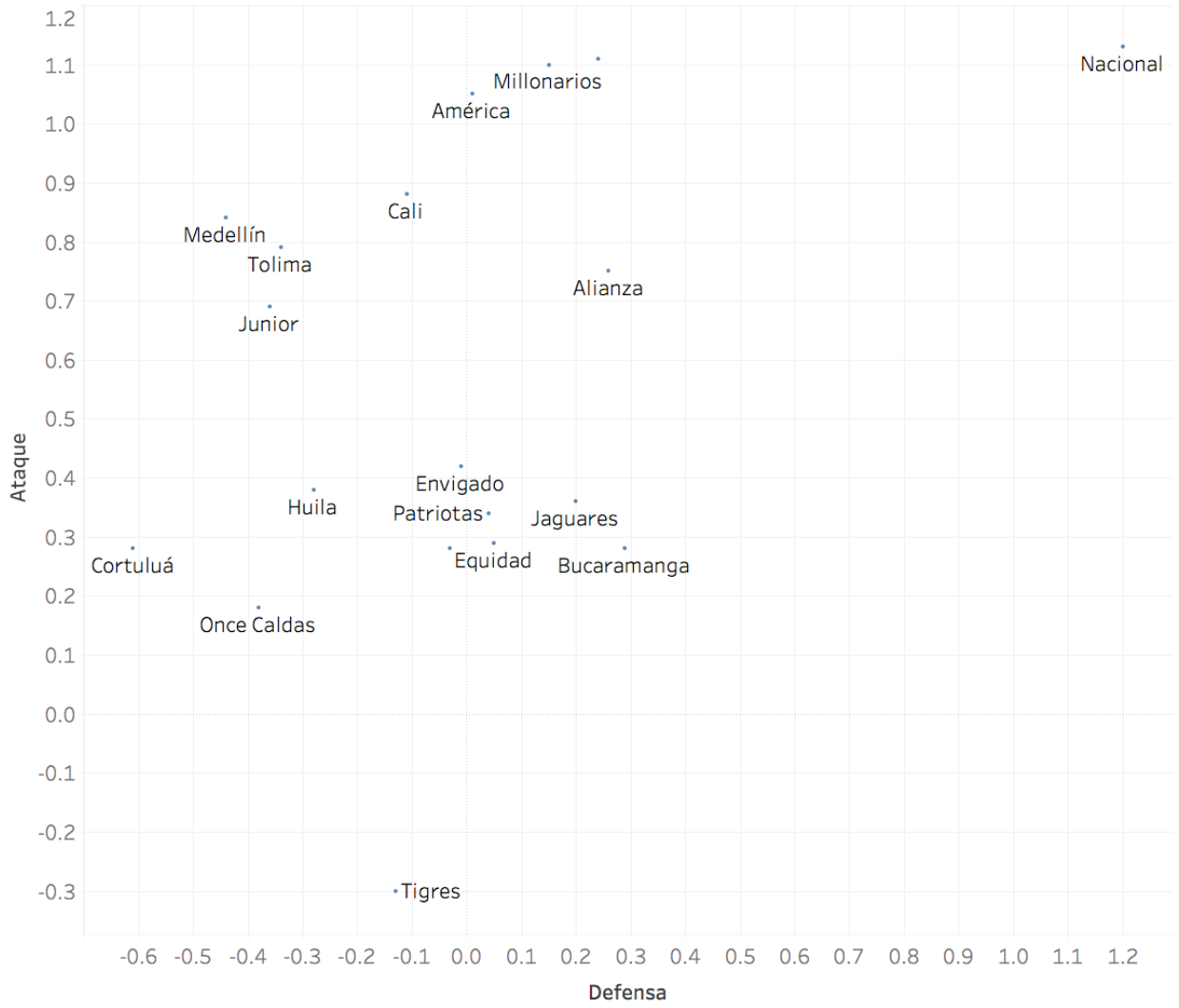

Gráfico 2. Visualización de los parámetros resultantes del modelo con datos de FIFA (parámetros defensivos se presentan con signos invertidos)

Conclusiones

En este artículo se explora la posibilidad de utilizar los datos derivados del videojuego FIFA como una herramienta que apoye la toma de decisiones en el fútbol, un deporte que – por sus características intrínsecas – anda rezagado frente a otros deportes en este rubro. En particular, se explora el uso de los datos de FIFA como una fuente de apoyo para la predicción de resultados de equipos en el fútbol, utilizando datos de la liga colombiana para el año 2017. Se encontró que la inclusión de los datos de FIFA mejoró la capacidad predictiva de los modelos planteados, un resultado que corrobora el potencial que podría tener esta fuente de datos como herramienta para

análisis cuantitativo de este deporte. En un futuro se espera explorar otras posibles aplicaciones de esta fuente de datos.

Referencias

Brechot, M., & Flepp, R. (2020). *Dealing With Randomness in Match Outcomes: How to Rethink Performance Evaluation in European Club Football Using Expected Goals. Journal of Sports Economics* 1-28

Burn-Murdoch, J. (2018) *How data analysis helps football clubs make better signings.* Financial Times. Tomado de: https://www.ft.com/content/84aa8b5e-c1a9-11e8-84cd-9e601db069b8

Cotta l., POSV de Melo, F Benevenuto, AAF Loureiro , (2016*) Using FIFA Soccer video game data for soccer analytics.* Workshop on Large Scale Sports Analytics

Fernández J., Bornn L., Cervone D. (2020) *A framework for the fine-grained evaluation of the instantaneous expected value of soccer possessions*

Freeman, L. (2019) *The Impact of Analytics Adoption on Team Performance in Professional Sports: A Longitudinal Analysis of the Lag in Observable Results.* 9 Proceedings of the Conference on Information System. Tomado de: http://proc.conisar.org/2019/pdf/5201.pdf

Karlis D., Ntzoufras I. (2003) *Analysis of sports data by using bivariate Poisson models.* The Statistician 52(3): 381 - 393

Liu, H. et al. (2015) *Match statistics related to winning in the group stage of 2014 Brazil FIFA World Cup.* Journal of Sports Science, 33:12, 1205-1213

Müller O., Simons A., Weinmann M. (2017) *Beyond crowd judgements: Data-driven estimation of market value in association football* Europoean Journal of Operational Research, 263(2):611–624

Paine N., Herring C. (2021) *James Harden's Rockets Changed The NBA Forever* FiveThrityEight. Tomado de: https://fivethirtyeight.com/features/james-hardens-rockets-changed-the-nba-forever/

Schoenfeld, B. (2019) *How Data (and Some Breathtaking Soccer) Brought Liverpool to the Cusp of Glory* . New York Times Magazine. Tomado de: https://www.nytimes.com/2019/05/22/magazine/soccer-data-liverpool.html

Smith R., (2016) *How Videogames Are Changing The Way Soccer is Played* New York Times. Tomado de: https://www.nytimes.com/2016/10/14/sports/soccer/the-scouting-tools-of-the-prosa-controller-and-a-video.html

Soto-Valero, C. (2017*). A Gaussian mixture clustering model for characterizing football players using the EA Sports' FIFA video game system*. RICYDE. Revista internacional de ciencias del deporte, 49(13), 244-259. https://doi.org/10.5232/ricyde2017.04904

Souza, D. B. (2019) *A new paradigm to understand success in professional football: analysis of match statistics in LaLiga for 8 complete seasons.* Internation Journal of Performance Analysis in Sport, DOI: 10.1080/24748668.2019.1632580

Verstraete K., Decroos T., Coussement B., Vannieuwenhoven N., Davis J. (2020*) Analyzing Soccer Players' Skill Ratings Over Time Using Tensor-Based Methods*. In: Cellier P., Driessens K. (eds) Machine Learning and Knowledge Discovery in Databases. ECML PKDD 2019. Communications in Computer and Information Science, vol 1168. Springer, Cham. https://doi.org/10.1007/978- 3-030-43887-617

Wolfe, R., Wright, P. M., & Smart, D. L. (2006). *Radical HRM innovation and competitive advantage: The Moneyball story.* Human Resource Management, 45(1), 111–145. doi:10.1002/hrm.20100