¿CUÁLES SON LOS DETERMINANTES DE LOS RESULTADOS EN EL FÚTBOL PROFESIONAL COLOMBIANO?

AUTORES:

GIOVANNY ALBERTO DE LA CRUZ ZULUAGA

NICOLAS CAMPO RAMÍREZ

DIRECTOR DEL PROYECTO:

PH.D. JULIO CÉSAR ALONSO / DIEGO ANTONIO BOHÓRQUEZ

UNIVERSIDAD ICESI

FACULTAD DE CIENCIAS ADMINISTRATIVAS Y ECONÓMICAS ECONOMÍA Y NEGOCIOS INTERNACIONALES

SANTIAGO DE CALI 2018

TABLA DE CONTENIDO

RESUMEN

Actualmente Colombia posee una de las ligas de futbol más competitivas del mundo, razón por la cual es motivo de investigación. Posterior a una revisión de literatura, se ha podido evidenciar la poca información que hay acerca de la liga colombiana y la escasez de estudios sobre la misma, siendo éste el primero en indagar sobre las variables más relevantes que influyen en la cantidad de goles anotados por un equipo profesional colombiano. En esa medida, el objetivo del presente estudio consiste en encontrar cuáles son los determinantes que inciden en el marcador final de un encuentro en la primera división del Futbol Profesional Colombiano. Ello mediante la implementación de un modelo de regresión de Poisson y la aplicación de criterios de información que contribuyan a especificar las variables que mejor explican al modelo. En el transcurso de la investigación se encontraron tres modelos y, mediante la aplicación de métricas de evaluación del poder predictivo, se seleccionó el mejor. Se logró determinar que variables como la altitud de la ciudad en donde se juega de local, no jugar en su estadio cuando juega de local, racha de victorias, promedio de goles en contra y jugar un partido de torneo internacional inmediatamente después del partido de liga inciden en el resultado final.

Palabras claves: Predicción, Modelo de regresión de Poisson, Fútbol, Liga Águila.

ABSTRACT

Currently Colombia has one of the most competitive football leagues in the world, which is why it is a matter of investigation. After a review of literature, it has been possible to show the little information that there is about the Colombian league and the scarcity of studies about it, being the first one to inquire about the most relevant variables that influence the amount of goals scored by a Colombian professional team. To that extent, the objective of this study is to find out what are the determinants that affect the final score of a match in the first division of Colombian Professional Soccer. This is done through the implementation of a Poisson regression model and the application of information criteria that help to specify the variables that best explain the model. During the course of the investigation, three models were found and, by applying predictive power evaluation metrics, the best one was selected. It was possible to determine that variables such as the altitude of the city where the team play at home, not play at their stadium when the team play at home, winning streak, average goals against and playing an international tournament match immediately after the league match affect the final result.

Keywords: Forecasting, Poisson regression model, Soccer, Liga Aguila.

1. INTRODUCCIÓN

El fútbol en Colombia es importante. Esto queda demostrado en el Plan Decenal de Fútbol realizado por el gobierno de Juan Manuel Santos, que muestra que el 94% de los colombianos considera al fútbol como trascendente en la sociedad del país. Se reconoce al fútbol recreativo como una "*actividad pedagógica que permite formar a las personas en valores como la disciplina, la solidaridad, ciudadanía, convivencia, tolerancia y respeto por el otro, y que puede ser aplicado indistintamente en todos los estratos socio-económicos, etnias y sexos".*[1] Por lo tanto, dado el interés que existe sobre el fútbol, cobra importancia estudiar y caracterizar, a través de los datos, el desarrollo de la liga colombiana, para poder tener un mejor entendimiento de su presente y poder comparar con otras ligas. Por tanto, el objetivo de esta investigación es encontrar los determinantes que inciden en el resultado final de los partidos de la primera división del fútbol profesional colombiano, de manera que sea posible contribuir al entendimiento de las dinámicas de dicho torneo.

Los datos de las variables que aquí se emplean para la estimación, son una recopilación de la temporada 2015 – I hasta la temporada 2018 – I para la liga colombiana. Dentro de las fuentes de recolección para la base de datos se destacan la página oficial de la Dimayor, diario AS y WinSports.

Por otra parte, en cuanto a la metodología empleada, se determinó dividirla en tres partes. La primera consiste en determinar qué modelo se ajusta mejor a las pretensiones de la investigación. Luego, se explica el método para la selección de las variables mediante dos muestras de entrenamiento constituidas con diferentes conjuntos de variables. Posterior a

[1] https://www.mininterior.gov.co/el-poder-del-futbol-la-gran-encuesta

ello, se pretende aplicar a cada una de estas muestras un criterio de información diferente, en este caso, Akaike Information Criterion (AIC) y Bayesian Information Criterion (BIC), a fin de evaluar cuáles son las variables que mejor se ajustan al modelo. Por último, se evalúa la capacidad predictiva del modelo contrastándolo con una muestra de evaluación y bajo dos métricas, las cuales arrojarán la probabilidad de acierto del modelo.

El presente documento consta de seis secciones adicionales. En la segunda sección se muestran los diferentes trabajos con los cuales se fundamenta el presente, es una breve revisión de la literatura a el fin de dar un sustento teórico a la investigación. En la tercera sección se realiza la presentación de los datos y las correspondientes estadísticas descriptivas. En la cuarta, se presentará la metodología utilizada para el estudio. La quinta sección evidencia los resultados obtenidos a lo largo de la investigación. En la última sección, se presentan algunas conclusiones y comentarios finales.

2. REVISIÓN DE LA LITERATURA

En esta sección se presenta una serie de estudios que serán la base teórica para el desarrollo del trabajo. Comenzando con Maher (1982) en su artículo "Modelling association football scores", quien plantea un modelo que tiene en cuenta parámetros de ataque y defensa para cada equipo y el efecto de jugar de local, de manera que sea posible modelar los goles anotados por un equipo de fútbol en un partido. Para ello, utiliza distribuciones de Poisson independientes y luego compara las frecuencias mediante pruebas de bondad de ajuste. Concluye que, un modelo independiente de Poisson ofrece una descripción razonablemente precisa para los goles anotados en un partido por un equipo, incluso más que un modelo con distribución binomial negativa. Adicionalmente deja abierta la discusión, proponiendo que una Poisson Bivariada también podría ajustarse bien.

Por otro lado, Dixon & Coles (1997) motivados por explotar las ineficiencias potenciales en el mercado de apuestas deportivas, parten del Modelo Poisson de Maher (1982) con el objetivo de ajustar mejor el resultado del modelo en partidos donde el número de goles sea bajo. Desarrollan un modelo basado en los datos de la liga inglesa desde 1992 a 1995. Al mismo tiempo los parámetros de ataque y defensa se vuelven dinámicos, es decir que dicha cuantificación está basada en el rendimiento reciente del equipo. Concluyen que el modelo tiene un rendimiento positivo cuando se utiliza como base de una estrategia de apuestas.

Dyte & Clarke (2000) sugieren un modelo cuyo objetivo es predecir la distribución de las anotaciones en los partidos de fútbol internacionales. Tratan los goles de cada equipo como variables con una distribución de Poisson independientes cuya media condicionada depende de la clasificación de cada equipo de la Federación Internacional de Fútbol (FIFA) y la condición del equipo en el partido (local, visitante o neutral). Los resultados de la regresión

de Poisson de este modelo se utilizaron para simular los partidos jugados durante el torneo de la Copa Mundial de 1998. Concluyen que el mejor modelo para predecir resultados debe seguir una distribución Poisson y que para que el modelo sea un predictor más efectivo, se deben hacer algunos ajustes manuales a los datos de calificación.

Por su parte, Goddard & Asimakopoulos (2002) elaboraron una obra en la cual modelaron resultados de partidos de fútbol, investigación titulada "Modelling football match results and the efficiency of fixed-odds betting". Ellos utilizan un modelo de regresión Probit, con datos de 15 años, para modelar los resultados de los partidos de fútbol de la liga inglesa. Además de los datos de los resultados de los partidos anteriores, en este paper se resalta la importancia de los partidos correspondientes al final de la temporada. Concluyen que, los resultados pasados de cada equipo en los partidos de liga son un factor que pudiese determinar el marcador final. Adicionalmente, le dan una valoración relevante a la participación de los equipos en la competición de la copa local, la participación en certámenes internacionales y a la distancia geográfica entre los lugares del equipo local y el equipo visitante, pues consideran que éstas variables contribuyen al desempeño del modelo.

Pollard (2008) en su obra *Home Advantage in Football: A Current Review of an Unsolved Puzzle* se plantea como objetivo evidenciar la incidencia que tiene el factor local en el rendimiento de los equipos, pues en él se involucran otros componentes relacionados con el efecto que tiene el público en el estadio local. El autor menciona la repercusión que llega a tener su proximidad al terreno y el tamaño de la hinchada. Por otra parte, se refiere al efecto del viaje puesto que la distancia recorrida puede influir en el rendimiento del equipo, y a la predisposición del árbitro, ya que existe evidencia de que las decisiones arbitrales favorecen al equipo local, usualmente como consecuencia de la presión del público. Adicionalmente,

Pollard habla acerca de la existencia de un factor psicológico, consecuencia de que los jugadores y entrenadores tienen conocimiento del efecto de jugar en casa. Por lo tanto, su actitud antes y durante del encuentro se ve perjudicada. Mediante una exhaustiva revisión bibliográfica llega a la conclusión de que las causas precisas de la ventaja en casa y la forma en que operan todavía no están bien entendidas. Una revisión de la evidencia a favor y en contra debe establecerse en un contexto de los siguientes hechos básicos: La ventaja en el hogar ha existido al menos desde el inicio del fútbol organizado a fines del siglo XIX; es un fenómeno mundial, pero varía considerablemente de un país a otro; ha disminuido en las mayores ligas de Europa en los últimos 15 años; tiende a ser mayor en el fútbol que en otros deportes de equipo.

Otro factor presente en la literatura es el efecto de ser un equipo recién ascendido y de ser un equipo "grande". Oberstone (2009) explica el éxito relativo de los clubes de fútbol de la Premier League. Su estudio fue realizado con datos de la temporada 2007 – 2008 de la Premier League, donde, mediante un modelo de regresión múltiple, identificó aquellas acciones específicas que separan estadísticamente a los clubes en estos tres grupos. Mediante la elaboración de una ANOVA de una vía concluye que, existe claramente una diferencia en el rendimiento de los cuatro primeros clubes de la tabla, los tres últimos y los doce restantes de la mitad.

Gómez (2013), es el único trabajo realizado para el futbol profesional colombiano hasta el momento. En él se pretende obtener el pronóstico de los resultados de los próximos partidos en el futbol profesional colombiano (FPC) y saber qué equipos ingresarán a la fase final. La metodología econométrica utilizada para el pronóstico de resultados es por medio de un modelo Logit Ordenado. Encuentra que, por ejemplo, cuando el equipo juega de local, la

probabilidad de obtener la victoria es mayor, o que cuando el equipo obtiene goles en la fecha anterior, la probabilidad de que anote en el próximo partido es mayor.

Skinner & Freeman (2009) son otros dos autores que resaltan la importancia de un modelo que siga una distribución Poisson. El objetivo de este trabajo es cuantificar la probabilidad de que haya sido o no el caso de que "el mejor equipo ganó" el partido. Proceden considerando un partido de fútbol como un experimento para evaluar cuál de los dos equipos es superior y examinan la probabilidad de que el resultado del experimento (partido) realmente represente las habilidades relativas de los dos equipos. Concluyen que, para los marcadores típicos, la probabilidad de un resultado engañoso es significativa. Modificar las reglas del juego para aumentar el número típico de goles marcados mejoraría esta situación.

Koopman & Lit (2012) retoman el trabajo de Karlis y Ntzoufras (2003) en una investigación cuyo objetivo es diseñar un modelo de Poisson bivariado dinámico, generalizando el modelo planteado en 2003 a fin de agregar variables que cambian con el trascurrir del tiempo. El desempeño de la muestra de la metodología se verifica en una estrategia de apuestas que se aplica a los resultados de los partidos de las temporadas 2010/11 y 2011/12 de la Premier League inglesa. Este trabajo presenta una novedad en el análisis estadístico de series temporales de resultados de partidos de fútbol u otros deportes de equipo.

Otro trabajo que tiene en cuenta el factor temporal en la producción del número de goles es Malcata, Hopkins, & Richardson (2012). En este estudio se informa la progresión de la puntuación de tres equipos de una academia de desarrollo de talento juvenil durante cinco temporadas utilizando un novedoso enfoque analítico basado en modelos mixtos generalizados con distribución Poisson. Se encuentra que este modelo tiene una utilidad

marginal para estimar la progresión de los resultados de fútbol, debido a la incertidumbre derivada de los marcadores bajos del juego.

García (2014) por su parte se propone como objetivo modelar los resultados de los partidos de fútbol de la liga española mediante el uso de modelos de regresión de Poisson bivariado con datos análisis pertenecientes a las temporadas 2010/2011 y 2011/2012. Concluye que, la ventaja de jugar como local no es universal, sino que parece depender del equipo y sus jugadores y deja abierto su trabajo para que futuros trabajos estudien si los resultados de una temporada parecen estar influidos por la percepción que tiene cada equipo de la ventaja de jugar en casa. Es decir, dado que los resultados matemáticos reflejan que existe una ventaja de jugar como local, sería para ella interesante observar si diferentes variables psicológicas pueden explicar esas percepciones de jugar de forma diferente debido a la circunstancia de ser equipo local o visitante.

3. PRESENTACIÓN DE LOS DATOS

Para llevar a cabo el presente estudio, se utilizaron como fuente 1488 partidos de fútbol, que se disputaron en la primera división del Futbol Profesional Colombiano, desde la temporada 2015 – I hasta el año 2018 – I. Los datos necesarios para la investigación provienen de Dimayor.com.co, AS.com, Winsports.com.

Teniendo como base la revisión de literatura y el deseo de probar la incidencia que tienen otras variables en el resultado de un partido de la liga, se utilizarán las siguientes variables continuas y dicotómicas. Así pues, las variables continuas fueron: PTOS.ACUM que reporta los puntos acumulados de cada equipo antes del partido a disputar, ALTITUD reporta la altitud del estadio donde juega de local, R.VIC es la cantidad de partidos que lleva ganados consecutivamente, R.DER es la cantidad de partidos que lleva siendo derrotado consecutivamente, R.INV es la cantidad de partidos que se ha mantenido sin perder de manera consecutiva, PROM.GF es el promedio de goles a favor y PROM.GC es el promedio de goles en contra.

Las variables dicotómicas fueron: LOCAL, que toma el valor de uno si el equipo jugó el partido en condición de local y cero en caso contrario; REC.ASCEND toma el valor de uno si el equipo recién ascendió a la primera división en la temporada del partido a disputar, y cero en caso de que no lo sea; NJESUESTDELOC toma el valor de uno si el equipo no jugó en su estadio cuando se presentó como local y cero en caso contrario; FAS.FIN es uno si el partido pertenece a la fase final del torneo o cero en caso de que no lo sea; T.INTER.ANTES toma el valor de uno si jugó un partido internacional antes del partido de liga y cero si no fue así; T.INTER.DESPUES toma el valor de uno si jugó un partido internacional después del partido liga y cero si no fue así; CAMBIODT es uno si el equipo cambió de director técnico

o cero en caso contrario; y por último, una variable dummy por cada equipo que participó en los torneos que hacen parte de la muestra, dejando por fuera al club Atlético Nacional, el cual es el equipo tomado como base para el estudio.

A continuación, se presentan dos tablas con información relevante de las variables que serán utilizadas en el estudio. La Tabla 1 muestra las estadísticas descriptivas de las variables continúas y la Tabla 2 muestra la frecuencia relativa de las variables dicotómicas, exceptuando a las variables dummy que representan a cada equipo participante.

Tabla 1. Estadísticas descriptivas.

Variable	**Media (Desviación estándar)**
PTOS.ACUM	14,46 (10,87161)
ALTITUD	1543 (958,4249)
R.VIC	0,546 (0,9634834)
R.DER	0,501 (0,8599388)
R.INV	1,427 (2,153935)
PROM.GF	0,9185 (0,4548001)
PROM.GC	0,883 (0,4327771)

Fuente: Elaboración propia.

Tabla 2. Frecuencia relativa de las variables dicotómicas.

Variable	**Frec. Rel. (%)**
LOCAL	50%
REC.ASCEND	11,09%
NJESUESTDELOC	3,43%
FAS.FIN	6,59%
T.INTER.ANTES	4,50%
T.INTER.DESPUES	4,84%
CAMBIODT	1,34%

Fuente: Elaboración propia.

4. METODOLOGIA

El primer paso de la metodología consiste en determinar el modelo que se va a emplear para las predicciones. De esta manera, teniendo en cuenta la revisión de la literatura presentada y dado que los goles marcados por un equipo en un partido son una variable de conteo, para este estudio se asume una distribución de Poisson independiente. El paso siguiente es explicar el método para la selección de las variables mediante una muestra de entrenamiento y el uso de criterios de información. Por último, en esta sección se pretende explicar las métricas empleadas para la elección del mejor modelo. Así entonces, se procede a explicar la metodología.

4.1. ESTABLECER EL MODELO:

Como ya se mencionó, se ha determinado seguir la metodología propuesta por la literatura empleando un modelo con distribución de Poisson. Este modelo de regresión es un tipo de Modelo Lineal Generalizado, lLos cuales permiten incluir distintas relaciones entre las medias condicionales de las variables respuesta y las explicativas. El modelo de regresión de Poisson se utiliza para datos de conteo como, por ejemplo, los goles que marca un equipo en un partido.

En el modelo de regresión de Poisson la media (λ) se explica en términos de las variables explicativas mediante el uso de un enlace. La función de probabilidad de una variable con distribución Poisson se modela de la siguiente forma:

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

Donde λ es el parámetro de la distribución o el número de veces que se espera que ocurra el evento en un período determinado, que es igual a la media y a la varianza. Por otro lado, *x* representa el número de veces que ocurre un evento de interés.

Ahora bien, habiendo mencionado que se asume que los goles siguen una distribución de Poisson, entonces se debe expresar lo siguiente:

$$g_i \sim Poisson\ (\ \lambda_i\ )$$

Donde g_i son los goles anotados por el equipo i, y λ_i es el valor esperado de los goles. Es decir:

$$E[g_i] = \lambda_i$$

Como la variable dependiente nunca tomará valores negativos, no es posible utilizar un modelo lineal directo y, por tanto, se necesita una función de enlace. Adicionalmente, es importante mencionar que, en la distribución de goles, la función de regresión está en el intervalo (0, +∞). Por tanto, el logaritmo parece la función de enlace más adecuada. Se expresaría de la siguiente forma:

$$\log(\lambda_i) = X\beta$$

O lo que es igual, la función de regresión del modelo de Poisson se formularía de la siguiente manera:

$$\lambda_i\ = \exp(X\beta)$$

Mediante $X\beta$ se representa el producto escalar del vector de variables explicativas por el vector de parámetros. Teniendo en cuenta lo anterior, se tiene entonces que los goles que marca un equipo siguen una distribución Poisson y se determinan de la siguiente forma:

$$g_i \sim Poisson(\exp(X\beta))$$

Por último, para estimar los parámetros del modelo se utiliza la función de máxima verosimilitud. Además, es conocido que los estimadores de máxima verosimilitud son asintóticamente normales y centrados y su matriz de varianzas-covarianzas es la inversa de la matriz de información (la matriz hessiana cambiada de signo), lo cual permite hacer inferencias sobre los parámetros del modelo (García, 2014).

Las estimaciones se llevan a cabo en el Software R, donde los modelos lineales generalizados se ajustan con la función *glm* y en la regresión de Poisson se debe especificar *family=poisson(link=log).*

4.2. MÉTODO DE SELECCIÓN DE VARIABLES:

Para este punto, la selección de las variables se lleva a cabo mediante una muestra de entrenamiento. Dicha muestra de entrenamiento posee información perteneciente a las temporadas 2015 – I a 2017 – II. Los datos de la temporada 2018-I serán utilizados como muestra de validación para contrastar los resultados del modelo. Básicamente lo que se pretende es pronosticar los resultados de los partidos en 2018-I (en la tercera etapa de la metodología) para evaluar la capacidad predictiva del modelo.

Ahora bien, en cuanto a la elaboración de la muestra de entrenamiento se conocía la existencia de una fuerte correlación entre las variables "racha de invicto" y el conjunto "racha de victorias" y "racha de derrotas". Por lo tanto, fue necesario separar dichas variables conformando dos subgrupos.

El primero contiene todo el conjunto de variables excluyendo la racha de victorias, mientras que, el segundo, posee todas las variables a excepción de la racha de victorias y la racha de derrotas.

Posteriormente, se procede a seleccionar las variables relevantes, al estimar modelos de regresión de Poisson, empleando los dos subgrupos de variables y decidiendo mediante los criterios de información AIC y BIC. Estos dos criterios son los más comunes en la literatura, y la principal razón que motiva a la utilización de estos criterios es que la estimación de factores por el método de máxima verosimilitud tiende, por lo general, a elegir un número mayor de factores, es decir, a elegir un modelo con exceso de parámetros (Caballero, 2011).

El Akaike Information Criterion (AIC) penaliza un exceso de parámetros ajustados, algo que no hace el test asintótico de la chi-cuadrado. Así entonces, se selecciona el modelo que mejor se ajusta a los datos, garantizando que es el mejor entre los modelos candidatos, en el sentido de que proporciona la aproximación más cercana a la realidad. La expresión matemática viene dada por la siguiente expresión general:

$$AIC(k) = -2 \ln \sigma\left[\hat{\beta}(k)\right] + 2k$$

En donde $\sigma\left[\hat{\beta}(k)\right]$ es la función de verosimilitud de las observaciones, $\hat{\beta}(k)$ es la estimación de máximo verosimilitud del vector de parámetros β y k es el número de parámetros independientes estimados dentro del modelo.

Por su parte, el Bayesian Information Criterion (BIC), a diferencia del AIC, considera el número de observaciones en el término de penalización, y es además menos favorable a la inclusión de varaibles. Su expresión matemática se define como:

$$BIC(k) = -2 \ln \sigma\left[\hat{\beta}(k)\right] + (\ln n)k$$

En donde $\sigma\left[\hat{\beta}(k)\right]$ es la función de verosimilitud de las observaciones, $\hat{\beta}(k)$ es la estimación de máximo verosímil del vector de parámetros β y k es el número de parámetros independientes estimados dentro del modelo, mientras n es el tamaño de la muestra.

En el proceso de selección de variables, se emplea el algoritmo "Stepwise" del Software R, el cual cuenta con tres aproximaciones: El primero, es el Forward o "selección hacia adelante", el cual parte de un modelo vacío y se van agregando variables hasta que se obtiene el mejor valor del criterio de información (AIC o BIC); el segundo es el Backward o "eliminación hacia atrás", el cual parte de un modelo con todas las variables del estudio, pero en cada iteración se eliminan variables hasta que se encuentra el mejor valor del criterio de información; el tercero es el Both o "ambos" ("mixto"), el cual es una combinación de los dos anteriores. Dado que son dos conjuntos de variables, mediante este algoritmo se encuentran 4 modelos diferentes, dos con el criterio de información AIC y dos con el BIC.

4.3 MÉTRICAS DE EVALUACIÓN

Con los 4 modelos encontrados anteriormente, el objetivo en este punto es evaluar su poder predictivo, pronosticando los resultados de la temporada 2018-I y comparando con los

resultados observados. Para llevar a cabo la comparación, se decide adoptar la metodologia de evaluación propuesta por Schauberger & Groll (2018). Estos autores utilizan dos métricas:

- Tasa de aciertos: indica el porcentaje de aciertos obtenidos por el modelo. Un gran valor de esta tasa de clasificación refleja un buen ajuste. Matemáticamente se expresa de la siguiente manera:

$$Tasa\ de\ aciertos = \frac{Número\ de\ partidos\ acertados}{Total\ de\ partidos}\ x100\%$$

- Probabilidad Multinomial: es el promedio de las probabilidades de todos aquellos eventos que realmente ocurrieron. Entre mayor sea este porcentaje, mayor nivel predictivo tiene el modelo, pues estará asignando una mayor probabilidad de ocurrencia a los eventos que efectivamente suceden. Se define como:

$$Probabilidad\ Multinomial = \frac{\sum_{i=1}^{n} \hat{p}_i}{n}$$

  Donde $\hat{p}_i$ es la probabilidad que el modelo había determinado para el resultado que realmente ocurrió y, n, es el número total de partidos.

5. RESULTADOS

Los resultados de la elección del modelo que mejor describe el marcado final de los partidos de la primera división del fútbol profesional colombiano, se muestran en la Tabla 3. Es pertinente aclarar que el Modelo 2 seleccionado por BIC es idéntico al Modelo 1 elegido por BIC, por lo tanto, el primero no se muestra en la tabla.

Tabla 3. Modelos estadísticos.

# de Modelo	Modelo	CR_IN	PROB_IN	CR_OS	PROB_OS
1	Modelo 1 por AIC	51,09%	40,04%	54,90%	41,67%
2	Modelo 1 por BIC	48,44%	38,22%	54,90%	39,51%
3	Modelo 2 por AIC	50,93%	40,03%	54,90%	41,45%
4	Aleatorio			33,33%	33,33%

Fuente: Elaboración propia.

Luego de poner a prueba los modelos bajo los criterios anteriormente mencionados en la metodología, la tasa de acierto muestra que todos los modelos cuentan con la misma probabilidad de acierto. Sin embargo, teniendo en cuenta la tasa de acierto en la muestra de entrenamiento, el modelo que mayor probabilidad de acertar el marcador final de un partido de la primera división del fútbol profesional colombiano es el Modelo 1 por AIC. Además, esto se puede evidenciar bajo los otros criterios de selección.

Así pues, las variables significativas del modelo seleccionado para el estudio se muestran en la Tabla 4.

Tabla 4. Variables del modelo seleccionado para el estudio.

Variable	Coeficiente	Nivel de significancia
LOCAL	0,4542 (0,03461)	***
ALTITUD.A	0,000089 (0,00004699)	*
NJESUESTDELOC.A	-0,20501 (0,09653)	*
R.VIC.B	0,07697 (0,03618)	*
PROM.GC.B	0,10287 (0,043336)	*
T.INTER.DESPUES.B	0,22007 (0,087255)	*
(R.VIC.A)2	-0,006816 (0,0038842)	*
(R.VIC.B)2	-0,02095 (0,009595)	*
MEDELLIN.B	0,29104 (0,09348)	**
AMERICA.A	-0,29917 (0,12622)	*
AMERICA.B	0,30341 (0,12887)	*
CALI.B	0,39529 (0,097345)	***
MILLONARIOS.A	-0,18507 (0,10423)	*
MILLONARIOS.B	0,20366 (0,09824)	*
SANTAFE.A	-0,33638 (0,1073)	**
JUNIOR.B	0,2717 (0,097513)	**
ALIANZAPETROLERA.A	-0,25566 (0,09393)	**
ALIANZAPETROLERA.B	0,45735 (0,10318)	***
BUCARAMANGA.A	-0,27673 (0,10547)	**

BUCARAMANGA.B	0,36364 (0,11996)	**
ATL.HUILA.A	-0,32608 (0,090952)	***
ATL.HUILA.B	0,39869 (0,10598)	***
BOY.CHICO.A	-0,78892 (0,13949)	***
BOY.CHICO.B	0,65616 (0,10445)	***
CORTULUA.A	-0,24325 (0,093113)	**
CORTULUA.B	0,52788 (0,10017)	***
CUCUTA.A	-0,54565 (0,18514)	**
CUCUTA.B	0,76199 (0,13463)	***
DEP.TOLIMA.A	-0,20571 (0,07572)	**
DEP.TOLIMA.B	0,24368 (0,095592)	*
DEP.PASTO.A	-0,48517 (0,11495)	***
DEP.PASTO.B	0,53839 (0,10466)	***
ENVIGADOFC.A	-0,44007 (0,096182)	***
ENVIGADOFC.B	0,30768 (0,10159)	**
FORTALEZA.A	-0,61918 (0,17653)	***
FORTALEZA.B	0,66659 (0,12816)	***
JAGUARES.A	-0,35145 (0,096945)	***
JAGUARES.B	0,50364 (0,09729)	***
LAEQUIDAD.A	-0,48618 (0,11408)	***

LAEQUIDAD.B	0,37754 (0,10560)	***
LEONES.A	-0,79654 (0,27373)	**
LEONES.B	0,44031 (0,13600)	**
ONCECALDAS.A	-0,31542 (0,098680)	**
ONCECALDAS.B	0,47618 (0,097546)	***
PATRIOTAS.A	-0,56055 (0,11938)	***
PATRIOTAS.B	0,27224 (0,10911)	*
RIONEGRO.A	-0,46146 (0,10247)	***
RIONEGRO.B	0,34496 (0,09978)	***
TIGRES.A	-1,0088 (0,18868)	***
TIGRES.B	0,32999 (0,17421)	*
UNIAUTONOMA.A	-0,56637 (0,18468)	**
UNIAUTONOMA.B	0,55662 (0,15132)	***
CONSTANTE	-0,44670 (0,094153)	***

Niveles de significancia: 10% (*), 5% (**), 1(***).

Errores Estándar entre paréntesis.

Fuente: Elaboración propia.

Es importante destacar que el modelo presenta problemas de heteroscedasticidad, por lo que fue necesario corregirlo por medio de la estimación consistente de White.

Ahora bien, en lo que se refiere a las estimaciones, la Tabla 4 muestra que ser local, la altitud del estadio donde juega el local, el promedio de goles en contra del visitante y que el equipo

visitante tenga que jugar después un partido de torneo internacional, aumentan la cantidad de goles a favor. Por el contrario, no jugar en su estadio cuando el equipo está en condición de local, disminuye la cantidad de goles en un partido.

Por otro lado, se encuentra que, a mayor racha de victorias del equipo, menor es la cantidad de goles que puede marcar dicho equipo en un partido, mientras que a mayor racha de victorias del rival (después de 2 partidos en línea), menor la cantidad de goles que se le puede marcar. Finalmente, las estimaciones de la Tabla 4 demuestran que el club Atlético Nacional es el equipo con los mejores parámetros de ataque y defensa de la Liga Águila.

6. CONCLUSIONES Y RECOMENDACIONES

El fútbol en Colombia es importante, y dado el interés que existe sobre este, se planteó encontrar el modelo de predicción, que de mejor manera describiera los resultados y caracterizara el desarrollo de la liga colombiana. Con el objetivo de encontrar los determinantes que inciden en el resultado final de los partidos de la primera división del fútbol profesional colombiano y teniendo en cuenta la revisión de literatura, se utilizaron diferentes modelos de Poisson, escogiendo el modelo que tuviera una mayor tasa de acierto en el pronóstico de resultados.

Los resultados obtenidos en el proyecto permitieron llegar a conclusiones interesantes. En primer lugar, se evidenció que variables como la altitud de la ciudad en donde se juega de local, no jugar en su estadio cuando juega de local, racha de victorias, promedio de goles en contra y jugar un partido de torneo internacional inmediatamente después del partido de liga, inciden en el resultado final de los partidos de la primera división del fútbol profesional colombiano.

En segundo lugar, tal como plantea la revisión de la literatura, jugar en condición local es una variable que aumenta sustancialmente la probabilidad de ganar un partido de la Liga Águila. Por tanto, es posible afirmar que la liga colombiana sigue la tendencia de las ligas de fútbol internacional, en donde igualmente la condición de jugar de local pesa fuertemente en el resultado final de un partido.

En tercer lugar, se logró comprobar que el club Atlético Nacional tiene los mejores parámetros de ataque y defensa en la liga colombiana. Sin embargo, es importante destacar que el Deportivo Independiente Medellín, Deportivo Cali, Independiente Santa Fe y Junior,

tiene parámetros muy similares, por lo que los resultados muestran que las diferencias no son significativas.

Finalmente, recomendamos para estudios posteriores continuar refinando el modelo, agregando variables para las que actualmente no se tiene suficiente información. Asimismo, sería un buen ejercicio emplear diferentes métodos de estimación, tales como: random forest, redes neuronales o modelos de elección discreta multinomial.

7. BIBLIOGRAFÍA

Caballero, F. (2011). *Selección de modelos mediante criterios de información en el ánalisis factorial. Aespectos teóricos y computacionales.* Granada: Universidad de Granada.

Clarke, S., & Norman, J. (1995). *Home ground advantage of individual.*

Dixon, M. J., & Coles, S. G. (1997). *Modelling Association Football Scores and Inefficiencies in the Football Betting Market.* Lancaster: Journal of the Royal Statistical Society Series C Applied Statistics.

Dixon, M. J., & Robinson, M. E. (1998). *A birth process model for association.* The Statistician.

Dyte, D., & Clarke, S. R. (2000). *A Ratings Based Poisson Model for World Cup Soccer Simulation.* hawthorn: The Journal of the Operational Research Society.

Fabián, A. C. (2012). *Predicción de resultados de eventos deportivos:*. Madrid: Universidad Carlos III de Madrid. Departamento de Teoría de la Señal y Comunicaciones.

García, E. (2014). *Aplicación de Modelos de Regresión de Poisson Bivariados a los resultados de los partidos de la Liga Española de Fútbol.* Universidad de Vigo.

Goddard, J., & Asimakopoulos, I. (2002). *Modelling football match results and the efficiency of fixed-odds betting* . Swansea: University of Wales Swansea.

Gómez, C. (2013). *Modelo de predicción de resultados en el Futbol Profesional Colombiano.* Bogotá, DC: Universidad de la Sabana.

Karlis, D., & Ntzoufras, I. (2005). *Bivariate Poisson and Diagonal Inflated Bivariate Poisson Regression Models in R.* Atenas: JSS Journal of Statistical Software.

Koopman, S. J., & Lit, R. (2012). *A Dynamic Bivariate Poisson Model for Analysing and Forecasting Match Results in the English Premier League.* Tinbergen Institute Discussion Paper.

Leeds, M. A., & Leeds, E. (2009). *International Soccer Success and National Institutions.* Journal of Sports Economics.

Maher, M. (1982). *Modelling association football scores.* Sheffield: Sheffield University.

Malcata, R., Hopkins, W., & Richardson, S. (2012). *Modelling the Progression of Competitive Performance of an Academy's Soccer Teams.* Journal of Sports Science and Medicine.

MARTÍN, R. (2011). *Modelado de los resultados de partidos de fútbol y de las ineficiencias en los mercados de apuesta.* Leganés: UNIVERSIDAD CARLOS III DE MADRID.

Medina, O., & Ospino, N. (2018). *El ecosistema del fútbol en Colombia y su asociación con el direccionamiento Estratégico de los clubes en la primera división colombiana.* Bogotá, DC: Universidad del Rosario.

Oberstone, J. (2009). *Differentiating the Top English Premier League Football Clubs from the Rest of the Pack: Identifying the Keys to Success.* San Francisco: Journal of Quantitative Analysis in Sports.

Pollard, R. (2008). *Home Advantage in Football: A Current Review of an Unsolved Puzzle.* San Luis Obispo: The Open Sports Sciences Journal.

Schauberger, G., & Groll, A. (2018). *Predicting matches in international football tournaments with random forests.* Statistical Modelling.

Skinner, G., & Freeman, G. (2009). *Soccer matches as experiments: How often does the 'best' team win?* Journal of Applied Statistics.

Torres, A. (2012). *Análisis del mercadeo deportivo para el sector del fútbol en Colombia a partir de los casos más exitosos a nivel internacional.* Bogotá D.C: Pontifcia Universidad Javeriana.

Volf, P. (2009). *A random point process model for score in matches.* Journal Management Mathematics.