# Proyecto final. Predicción de resultados de fútbol

Martínez Hernández Dafne Angélica. angelicam16@ciencias.unam.mx
Olvera Trejo Alberto. alberto0410@ciencias.unam.mx

22 de enero de 2022

**Resumen:** En el presente proyecto se realizará un modelo matemático que permita predecir los resultados de una temporada de fútbol con base en los resultados de temporadas anteriores, haciendo uso de la distribución de Poisson y de distintos métodos vistos en el curso de Manejo de Datos.

**Palabras clave:** modelo, predicción, fútbol, distribucion Poisson.

## 1. Introducción

El predecir qué equipo va a ganar en un partido de football soccer siempre ha sido un tema que ha dado de qué hablar. En este trabajo se va a hacer uso de la distribución Poisson para crear un modelo que dada una base de datos, una temporada de entrenamiento y la temporada a predecir, despliegue como resultado qué equipo va a ganar.
La base de datos a manejar contiene la información relacionada con la primera y segunda división de fútbol de la liga Española de las temporadas 1970-71 - 2017-18, se hizo uso de datos como los nombres de los equipos locales y visitantes que jugaron los partidos, número de goles correspondientes a ellos en cada uno de los juegos, el ganador de cada partido (local, visitante o empate) y la temporada en la que se jugó cada uno de ellos.

## 2. Objetivos

- Obtener un modelo que prediga los resultados de una temporada de fútbol con base en alguna de las temporadas anteriores con un margen de precisión de entre el 40 % y el 50 %.
- Valorar el rendimiento de dicho modelo
- Crear visualizacion de los resultados

## 3. Marco teórico

La distribución de Poisson, siendo una distribución discreta, nos indica la probabilidad de que ocurra una cantidad determinada de eventos en un cierto periodo en tiempo; por lo que, tomando la duración de un partido de fútbol como el periodo de tiempo, la usaremos para calcular el número de goles marcados por el equipo local y el equipo visitante en cada uno de los partidos que se jugaron en dicha temporada, obteniendo así una predicción sobre qué equipo ganará dicho partido o si se obtiene un empate, basándose en los resultados de cada uno de los equipos en las temporadas anteriores a las que se busca predecir.

## 4. Metodología

Se creó una carpeta llamada ”Proyecto_final”que contiene los siguientes tres archivos tipo jupyter notebook:

- **funciones**.- En este archivo se almacenan todas las funciones a usar
- **procesamiento**.- En este archivo se manipula la base de datos para posteriormente poder aplicarle las funciones
- **main**.- En este archivo únicamente se ejecutan las funciones que muestran en pantalla los resultados del modelo

El primer paso fue utilizar *pandas* para poder cargar y limpiar la base de datos, por lo que se eliminaron las columnas que no se iban a utilizar y las restantes se renombraron para un mejor manejo. Posteriormente se transformo la columna de fechas a formato *datetime*. Por último se aplicó la función *ganador* a la base de datos,dicha función indica si el equipo local ganó, el visitante ganó o bien hubo un empate. La base de datos obtenida

se nombró con el nombre de *historico*.

| fecha | temporada | division | round | local | visitante | localGoles | vistanteGoles | ganador |
|---|---|---|---|---|---|---|---|---|
| 1970-12-09 | 1970-71 | 1 | 1 | Atletico de Bilbao | Barcelona | 1 | 1 | Empate |
| 1970-12-09 | 1970-71 | 1 | 1 | Las Palmas | Atletico de Madrid | 1 | 1 | Empate |

Figura 1: Base de datos *historico* después del procesamiento

Suponiendo que se quiere predecir la temporada $x$, entonces hay que tomar la temporada $x-1$ para poder entrenar el modelo. Así, de la base *historico* se seleccionan todos los partidos de la división 1 llevados acabo en la temporada $x$. Posteriormente se crean dos copias de dicho resultado, una de ellas será para los locales y otra para los visitantes. Una vez hecho, de la base *locales* se elimina la columna de *visitante* y hacemos se hace lo análogo para la base *visitantes*. Finalmente se renombran las columnas para que tengan más coherencia.

Posteriormente cada base de datos se agrupa por equipo para poder poder usar la función de agregación *mean* y así cada base contiene únicamente tres columnas, el nombre del equipo, el promedio de goles a favor y otra para el promedio de goles en contra, como se muestra en la imagen 2

| equipo | prom_gol_favor_local | prom_gol_cont_local |
|---|---|---|
| Almeria | 1.368421 | 1.631579 |
| Atletico de Bilbao | 2.210526 | 0.947368 |

Figura 2: Tabla locales_prom. La primer columna es el nombre del equipo, la segunda el promedio de goles que ese equipo anotó siendo local y la tercer columna es el promedio de goles que ese equipo recibió como local

Por ultimo se obtiene un promedio sobre el promedio, es decir, las nuevas columnas creadas (las de los promedios por equipo) se van a dividir respectivamente entre el promedio de los goles que metieron los locales en toda la temporada $x$, la de entrenamiento, y entre el promedio de los goles que metieron lo visitantes en la misma temporada $x$. Este ultimo valor es el factor de defensa (en el caso de los goles en contra) y el factor de ataque (en el caso de los goles a favor) de cada equipo.

Ahora que ya se tienen los factores de defensa y ataque de cada equipo en su papel de visitante y de local, lo que hace falta es juntar toda la información obtenida en una sola tabla pero en donde esten los datos de la temporada a predecir. Así, se toma el *historico* y se seleccionan todos los partidos de la temporada a predecir; posteriormente se cruza con la tabla de locales (la que contiene los factores de ataque y defensa de locales) y después a la tabla resultado se cruza con la de visitantes (con los factores de ataque y defensa de los visitantes). Después se eliminan todas las columnas que ya no se usan obteninedo unicamente los datos base (temporada, nombre del equipo vistante, local, el ganador) y los factores de defensa y ataque de cada uno en sus respectivos papeles de visitante o local

El siguiente paso es obtener los valores de lambda, los cuales se van a calcular como un promedio ponderado de la siguiente manera:

- Para los locales el valor de lambda es:

$$fac\_ataqueL + (fac\_defensaV)(promGolL), \tag{1}$$

donde *fac_ataqueL* es el factor de ataque del local, *fac_defensaV* el factor de defensa del visitante y *PromGolL* es el promedio de goles que meten los locales en la temporada $x$

- Para los visitante el valor de lambda es:

$$fac\_ataqueV + (fac\_defensaF)(promGolV), \tag{2}$$

donde *fac_ataqueV* es el factor de ataque del visitante, *fac_defensaL* el factor de defensa del local y *PromGolV* es el promedio de goles que meten los visitantes en la temporada $x$

Por ejemplo, para los locales, el valor de lambda es qué tanto meten goles sumado a qué tan bien defienden los visitantes multiplicado por el promedio de goles de los locales en la temporada $x$. Aquí es en donde el modelo toma en cuenta los valores de la temporada $x$.
Con las formulas 1 y 2 se crean dos columnas nuevas en la última tabla obtenida. El resultado ob-

tenido se llamará la tabla *lambdas*, la cual es la mostrada en la imagen 3

| | temporada | round | local | visitante | localGoles | vistanteGoles | ganador | fac_ataqueL_x | fac_defensaL_x | fac_ataqueV | fac_defensaV | lambdaL | lambdaV |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 2014-15 | 1 | Malaga | Atletico de Bilbao | 1 | 0 | Local | 0.833556 | 1.179635 | 0.833556 | 1.179635 | 2.621157 | 1.938819 |
| 1 | 2014-15 | 29 | Sevilla | Atletico de Bilbao | 2 | 0 | Local | 1.389260 | 1.179635 | 0.833556 | 1.179635 | 3.176861 | 1.938819 |

Figura 3: Tabla *lambdas*

Hay que recordar que se va a hacer uso de la función de distribución Poisson, la cual es:

$$\mathbb{P}(X = k) = \frac{e^{-\lambda}\lambda^k}{k!1}, \tag{3}$$

donde $X$ es la variable aleatorio que nos dice cuantos goles metió un equipo, $k$ es el número de goles y $\lambda$ es el valor ya calculado. Así, únicamente falta encontrar el rango sobre el cual va a correr $k$. Dicho problema se resuelve encontrar el máximo y mínimo número de goles que han metido los equipos locales en la base de datos *historico*, pues por el análisis exploratorio de datos se sabe que los locales meten más goles que los visitantes.

Una vez teniendo listos todos los datos necesarios para poder usar la función Poisson, se aplica en la tabla *lambdas* mediante un ciclo *for* que corre desde el mínimo de goles hasta el máximo de goles, obteniendo la tabla *pre_final* mostrada en la imagen 4: Después se calculó la probabilidad de

| temporada | round | local | visitante | localGoles | vistanteGoles | ganador | proba_0GolesL | proba_0GolesV | proba_1GolesL | ... |
|---|---|---|---|---|---|---|---|---|---|---|
| 2014-15 | 1 | Malaga | Atletico de Bilbao | 1 | 0 | Local | 0.072719 | 0.143874 | 0.190607 | ... |
| 2014-15 | 29 | Sevilla | Atletico de Bilbao | 2 | 0 | Local | 0.041716 | 0.143874 | 0.132527 | ... |

Figura 4: Base de datos *pre_final*

que ganara el local, la probabilidad de que ganara el visitante y la probabilidad de empate. El procedimiento para calcular la probabilidad de que el local ganara fue sumar todas las probabilidades en las que el número de goles del local era mayor que el de los visitantes; análogamente para la probabilidad de que el visitante ganara. Para el empate se sumaron las probabilidades de que se anotaran el mismo número de goles. Sólo resta asignar quién gana en cada partido, para ello se tomaron los siguientes criterios:

- Si la probabilidad de que gane el local es mayor que la probabilidad de que gane el visitante, entonces se asigna como ganador al local
- Si la probabilidad de que gane el visitante es mayor que la probabilidad de que gane el local, entonces se asigna como ganador al visitante
- Se asigna empate si la diferencia entre la probabilidad del local y la del visitante es muy chica (en este caso, menor que $0{,}05$). También se asigna empate si la probabilidad de empate es mayor que la del visitante y mayor que la del local

El primer criterio para el empate está basado en lo que haría un humano, pues si ambos equipos tiene (casi) la misma probabilidad de ganar, entonces se designa dicho resultado
Aplicando dichos criterios a la tabla *pre_final* de la imagen 4 se obtienen los resultados del modelo. Finalmente se eliminan las columnas que no se usarán y se renombra la tabla como *final*, la cual se muestra en la imagen 5

| temporada | round | local | visitante | localGoles | proba_local | proba_visitante | proba_empate | ganador | prediccion |
|---|---|---|---|---|---|---|---|---|---|
| 2014-15 | 1 | Malaga | Atletico de Bilbao | 1 | 0.615919 | 0.456463 | 0.184289 | Local | Local |
| 2014-15 | 29 | Sevilla | Atletico de Bilbao | 2 | 0.709989 | 0.461610 | 0.158264 | Local | Local |
| 2014-15 | 23 | Granada | Atletico de Bilbao | 0 | 0.572233 | 0.469595 | 0.193014 | Empate | Local |

Figura 5: Base de datos *final*

Para mostrar los resultados se creó la función $modelo(temp\_entr, temp\_predecir, historico, jornada = 0)$, la cual toma como parámetros de entrada obligatorios la temporada con la cual se va a entrenar el modelo, la temporada a predecir y la base de datos con los todos los partidos; y como parámetro opcional la jornada hasta la cual se quiere ver los resultados; si no se especifica el parámetro jornada, entonces se mostraran todas las jornadas. Dicha función regresa un análisis como se muestra en las figuras 6:, 7, 8

Como extra se decidió calcular la *precisión* y el *recall* del modelo con la métrica de evaluación, es decir, la *precisión* y el *recall* están dados por las formulas

$$Precision = \frac{TP}{TP + FP}, Recall = \frac{TP}{TP + FN}, \tag{4}$$

donde *TP* denota a los verdaderos positivos, *FP* a los falsos positivos y *FN* a los falsos negativos.

```
modelo('2015-16','2016-17', historico, 4)

***RESULTADOS TEMPORADA 2016-17

****Porcenaje de aciertos: 48.89705882352941%
***Top 5 equipos con más aciertos:

Para el equipo Atletico de Bilbao el modelo acertó a 23 partidos de 32
Para el equipo Las Palmas el modelo acertó a 21 partidos de 32
Para el equipo Deportivo el modelo acertó a 18 partidos de 32
Para el equipo Granada el modelo acertó a 18 partidos de 32
Para el equipo Barcelona el modelo acertó a 17 partidos de 32

***El porcentaje de aciertos sobre los locales es 65.19%
***El porcentaje de aciertos sobre los visitantes es 50.0%
***El porcentaje de aciertos sde empate es 20.48%
```

Figura 6: Resultados de la función *modelo*

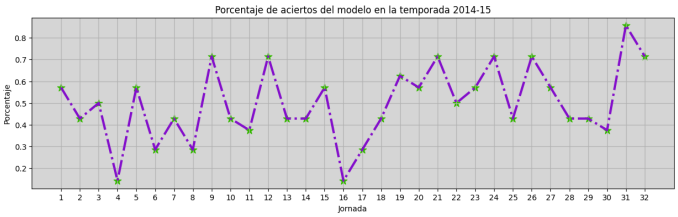


Figura 7: Gráfica de la función *modelo*. Porcentaje de aciertos por jornada en la temporada a predecir

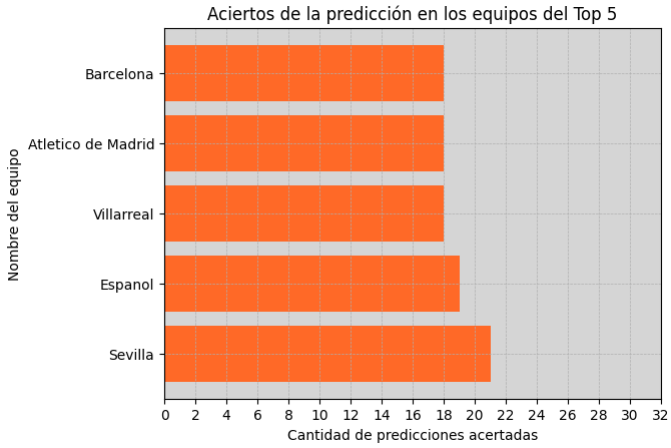


Figura 8: Gráfica de la función *modelo*. Top 5 de equipos a los cuales el modelo acertó más

Además, se observó que el equipo *Barcelona* apareció en 4 de los 5 top5 realizados, por lo que se analizó su comportamiento respecto a la predicción del modelo desde la temporada $2000-01$ hasta la $2016-17$

## 5. Resultados y discusión

Se aplicó el modelo con las temporadas 2011-12, 2012-13, 2013-14, 2014-15 y 2015-16 para predecir los resultados de las temporadas 2012-13, 2013-14, 2014-15, 2015-16 y 2017-18 respectivamente, obteniendo así una precisión de hasta el 49 % en la predicción los resultados de la temporada 2012-13.

El número máximo de aciertos en los resultadosn de los partidos jugados por un equipo en las temporadas mencionadas fue de 23, con los equipos de Atlético de Bilbao y Barcelona; siendo este último equipo aquel que apareció 4 de 5 veces en el "Top 5"de equipos que más predicciones acertadas tuvieron en una temporada. Tomando en cuenta dicha obserbación, se aplicó el modelo para poder realizar una gráfica que mostrara la cantidad de predicciones que el modelo acertó para el equipo Barcelona desde la temporada *2000-2001*.

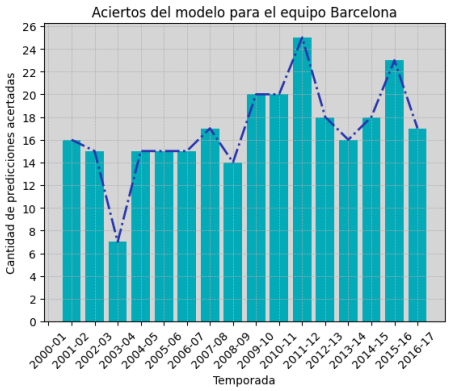


Figura 9: Cantidad de predicciones acertadas para el equipo Barcelona en cada temporada.

La precisión del modelo al predecir los partidos a favor de los equipos locales de la temporada 2012-13 a la 2016-17, es decir, de las veces que determinó que ocurriría cierto evento y ocurrió de esa manera, es del 60.23 %, a favor de los visitantes es del 24.54 % y de los empates de un 24.87 %.

El recall del modelo (de los eventos ocurridos cuántos de ellos se predijeron de manera correcta) es del 80.89 % para juegos a favor de los equipos locales, de un 3.21 % para los juegos a favor de los visitantes y de un 37.29 % para los empates.

# 6. Conclusión

El modelo cuenta con un considerable porcentaje de precisión, el cual tiene una mayor exactitud al predecir los resultados que serán a favor de los equipos locales y teniendo un gran déficit de ésta en la predicción de los resultados a favor de los equipos visitantes. Algo más que se pudo observar es que el equipo más *consistente* es el Barcelona, pues los en los partidos donde éste jugó, el modelo predijo correctamente una buena cantidad de los resultados de ellos.

# Referencias

[1] (2020). "Predicción en fútbol a partir de los goles marcados: la distribución de Poisson (I)". Consultado el 25 de enero de 2022. Obtenido de: https://vencex.com/2018/01/01/prediccion-en-futbol-a-partir-de-los-goles-marcados-la-distribucion-de-poisson-i/#:~:text=La%20distribuci%C3%B3n%20de%20Poisson%20es,marcados%20por%20el%20equipo%20X%C2%BB.

[2] . (2019). 'Clasificación con datos desbalanceados". Consultado el 25 de enero de 2022. Obtenido de: https://www.aprendemachinelearning.com/clasificacion-con-datos-desbalanceados/