Parcial 1

Juan Andrés Mejía Rocha

Procesamiento de datos a gran escala

Profesor Miguel Meléndez

Pontificia Universidad Javeriana
Bogotá D.C, Colombia

2022

Enunciado:

Supongan que ustedes son contactados por un inversor millonario que quiere crear un nuevo equipo de fútbol en la liga inglesa, él tiene el interés de conocer un poco más los equipos y resultados que se obtuvieron en la liga de 17-18, para así poder tener una mejor idea de que tipos de jugadores preferiría contratar y que estilo de juego quiere que tenga su equipo. Para esto ustedes cuentan con la información de los resultados por partido, como también las estadísticas obtenidas por los equipos y jugadores

Contexto del problema:

En el enunciado anteriormente descrito se nos plantea un contexto que requiere analizar el entorno de la liga de futbol inglesa (premier league) donde se quiere crear un nuevo equipo de futbol. Para la creación de este nuevo equipo se debe realizar un estudio de la información brindada de la liga, en ella hay datos y estadísticas de los diferentes jugadores y equipos. A partir de esta información se espera poder decir con certeza cuales son o serían los mejores jugadores para las posiciones de defensa, medio campo y delantera, además de una táctica en específico que el club deba seguir para tener la mayor posibilidad de ganar. La información para analizar sitúa el contexto en el año 2017 a 2018 donde el Manchester City fue el equipo ganador de la liga. Al final del documento de presenta un link donde se puede consultar la tabla del resultado de toda la liga.

Descripción de conjunto de datos

Para la realización de este análisis se proveen tres fuentes de datos de dos tipos de extensiones diferentes. Los archivos serán utilizados todos como dataframes en la plataforma de Google Collab

Players.csv:

El primer archivo es de extensión csv y es un listado de todos los jugadores de la liga organizados por sus equipos además de información adicional como su posición, edad y más.

	name	club	age	position	position_cat	market_value	page_views	fpl_value	fpl_sel	fpl_points	region	nationality	new_foreign	age_cat	club_id	big_club	new_signing
0	Alexis Sanchez	Arsenal	28	LW	1	65.0	4329	12.0	17.10%	264	3	Chile	0	4	1	1	0
1	Mesut Ozil	Arsenal	28	AM	1	50.0	4395	9.5	5.60%	167	2	Germany	0	4	1	1	0
2	Petr Cech	Arsenal	35	GK	4	7.0	1529	5.5	5.90%	134	2	Czech Republic	0	6	1	1	0
3	Theo Walcott	Arsenal	28	RW	1	20.0	2393	7.5	1.50%	122	1	England	0	4	1	1	0
4	Laurent Koscielny	Arsenal	31	CB	3	22.0	912	6.0	0.70%	121	2	France	0	4	1	1	0

De este listado, las variables más útiles pueden ser; de la posición donde juega el jugador, el club y el nombre:

- Posición
- Nombre
- Club
- Fpl_points

Resultados.csv:

Este segundo archivo de tipo csv muestran los resultados de los partidos de la Premier league durante el periodo del año 17 – 18. En este, se encuentra información detallada sobre cada partido, hora de inicio, los clubes participantes, ganador del partido además de las estadísticas del partido como la cantidad de off sites, corners, tiros al arco y más.

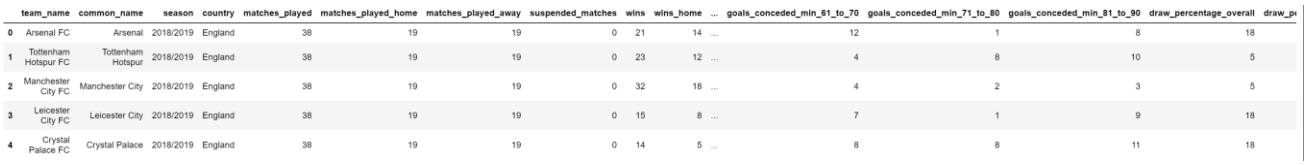
	Div	Date	HomeTeam	AwayTeam	FTHG	FTAG	FTR	HTHG	HTAG	HTR	...	HST	AST	HF	AF	HC	AC	HY	AY	HR	AR
0	E0	11/08/17	Arsenal	Leicester	4	3	H	2	2	D	...	10	3	9	12	9	4	0	1	0	0
1	E0	12/08/17	Brighton	Man City	0	2	A	0	0	D	...	2	4	6	9	3	10	0	2	0	0
2	E0	12/08/17	Chelsea	Burnley	2	3	A	0	3	A	...	6	5	16	11	8	5	3	3	2	0
3	E0	12/08/17	Crystal Palace	Huddersfield	0	3	A	0	2	A	...	4	6	7	19	12	9	1	3	0	0
4	E0	12/08/17	Everton	Stoke	1	0	H	1	0	H	...	4	1	13	10	6	7	1	1	0	0

Algunas de las variables que pueden resultar importantes de este conjunto de datos pueden ser:

- HST: disparos al arco por parte del equipo local
- AST: disparos al arco por parte del equipo visitante
- HomeTeam: nombre del equipo local
- AwayTeam: nombre del equipo visitante

Teams.csv

Por último, el conjunto de datos de equipos donde se encuentran indicadores y estadísticas importantes de cada equipo perteneciente a la primera división de la premier league. Algunos ejemplos de variables son el nombre del equipo, cantidad de partidos ganados, perdidos, estadísticas de partidos ganados como visitantes y locales, etc...



	team_name	common_name	season	country	matches_played	matches_played_home	matches_played_away	suspended_matches	wins	wins_home	...	goals_conceded_min_61_to_70	goals_conceded_min_71_to_80	goals_conceded_min_81_to_90	draw_percentage_overall	draw_p
0	Arsenal FC	Arsenal	2018/2019	England	38	19	19	0	21	14	...	12	1	8	18	
1	Tottenham Hotspur FC	Tottenham Hotspur	2018/2019	England	38	19	19	0	23	12	...	4	8	10	5	
2	Manchester City FC	Manchester City	2018/2019	England	38	19	19	0	32	18	...	4	2	3	5	
3	Leicester City FC	Leicester City	2018/2019	England	38	19	19	0	15	8	...	7	1	9	18	
4	Crystal Palace FC	Crystal Palace	2018/2019	England	38	19	19	0	14	5	...	8	8	11	18	

Algunas de las variables consideradas como pertinentes para las consultas a realizar son:

- Average_possession: Promedio de la posesión total en la liga
- Average_total_goals_per_match: Promedio de goles por Partido
- Position_cat: Número identificador de la posición donde 1 significa delantero, 2 medio campo, 3 defensa y 4 portero
- Clean_sheets: numero de partidos en donde no se recibió ni un gol
- goals_conceded_per_match: cantidad de goles en contra por partido (promedio)
-

Calidad de los datos:

Para poder llevar a cabo un análisis de la calidad de los datos de manera uniforme, se han definido tres diferentes dimensiones de calidad sobre las cuales serán calificados los respectivos conjuntos de datos:

- Vigencia: La vigencia de los datos indica que tan actualizada esta la información
- Consistencia: La consistencia mide la completitud de cada atributo del dataframe

- Unicidad: En la unicidad se evalúa el nivel de duplicación entre los datos

Gráficas y tablas:

Con el fin de realizar un análisis satisfactorio se proponen algunas preguntas que pueden ser respondidas a partir de las querys realizadas además de ciertos enfoques divididos por cada categoría de posición de jugador como la defensa, el medio campo y la delantera.

Posición	Descripción
Portería y defensa	• En el caso de la portería y la defensa se considera que el indicador de cantidad de goles recibidos es un factor que puede indicar la efectividad del arquero y la defensa. Para evitar sesgos, las dos posiciones serán examinadas al tiempo debido a que se podrían presentar casos donde un portero reciba muchos goles por múltiples errores en la defensa o que una buena defensa reciba goles por un bajo rendimiento del portero.

Medio campo	• Con el medio campo después de leer ciertos artículos, se define que la importancia del medio campo esta en movilizar el juego, abriendo espacios para dejar que los delanteros hagan su tarea. Para la parte defensiva es parecido ya que deben ayudar tanto en la recuperación de la pelota como en la salida de la pelota en el territorio del equipo. Por lo tanto, la variable de posesión grosso modo indica la capacidad de juego del medio campo ya que para tener el balón hay que moverse bien y hacer buenos pases
Delantera	• Los delanteros están encargados de recibir el balón y marcar puntos (meter gol) por lo cual un buen indicador de que tan bien están jugando es el de la cantidad de tiros al arco y su efectividad a la hora de hacerlo donde con la cantidad de tiros al arco se ve la constancia del jugador y con los goles la efectividad

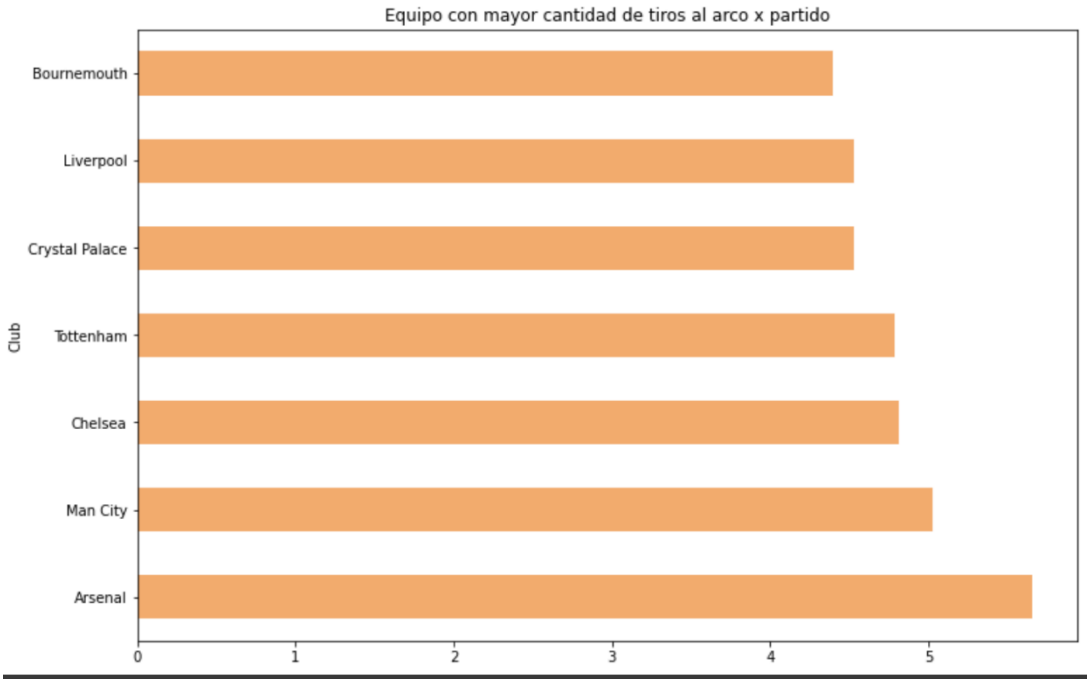
Algunas de las preguntas que se pueden plantear para responder son ¿A mayor posesión, más probabilidad de ganar y recibir menos goles? ¿Hay alguna forma de saber en medidas generales si los goles en contra son causados por parte de la defensa o de la portería? ¿El equipo que mas goles anota es el ganador de la liga?

Delantera

En esta tabla se observan los equipos con mayor promedio de goles por partido, equipos como el Manchester United, Manchester City, Arsenal y Liverpool se encuentran en un promedio de 3 goles por partido.

Arsenal FC	3.26
Manchester United FC	3.13
Manchester City FC	3.11
Fulham FC	3.03
Burnley FC	2.97
Watford FC	2.92
Liverpool FC	2.92

Por medio de esta grafica podemos ver como los equipos Arsenal, Manchester City, Chelsea y Tottenham, Crystal Palace y Liverpool



Equipo con mayor cantidad de tiros al arco:

Son aquellos equipos que más tiros al arco realizan, en base a estos datos se podría inferir que son las delanteras más efectivas para generar peligro y posibilidad de gol.

Listados de delanteros de los equipos con mayor cantidad de goles por promedio y mayores tiros al arco:

Alexis Sanchez	Arsenal	264	LW
Mesut Ozil	Arsenal	167	AM
Theo Walcott	Arsenal	122	RW
Olivier Giroud	Arsenal	116	CF
Alex Iwobi	Arsenal	89	LW

Roberto Firmino	Liverpool	180	SS
Philippe Coutinho	Liverpool	171	AM
Sadio Mane	Liverpool	156	LW
Adam Lallana	Liverpool	139	AM
Divock Origi	Liverpool	96	CF
Daniel Sturridge	Liverpool	54	CF
Ben Woodburn	Liverpool	5	LW
Sheyi Ojo	Liverpool	0	LW
Mohamed Salah	Liverpool	0	RW

Kevin De Bruyne	Manchester+City	199	AM
Sergio Aguero	Manchester+City	175	CF
Raheem Sterling	Manchester+City	149	LW
David Silva	Manchester+City	130	AM
Leroy Sane	Manchester+City	105	LW

Medio campo

En el medio campo como fue especificado anteriormente, la importancia de la posesión en el campo dado a que nos indica indirectamente que, para una mayor posesión, debe de haber buenos pases.

common_name	average_possession
Manchester City	68
Chelsea	64
Liverpool	62
Tottenham Hotspur	59
Arsenal	58

Esta tabla indica el valor por promedio del acumulado de todos los partidos de liga por equipo. Vemos que los equipos Manchester City, Chelsea y Liverpool tienen el 60% o más de la posesión total del balón en todos sus partidos.

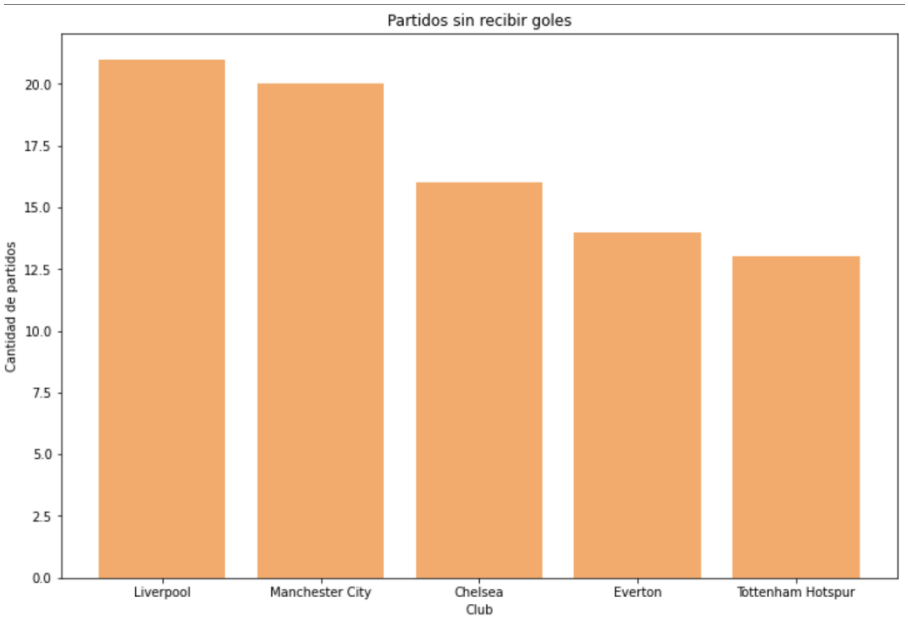
Jugadores interesantes para la posición pueden ser los mediocampistas del Manchester City o del Chelsea, a continuación, un listado de algunos de los jugadores con mayor puntuación en la liga de dichos equipos.

Cesc Fabregas	Chelsea	121	CM
Victor Moses	Chelsea	105	RM
Nemanja Matic	Chelsea	105	DM
N%27Golo Kante	Chelsea	83	DM

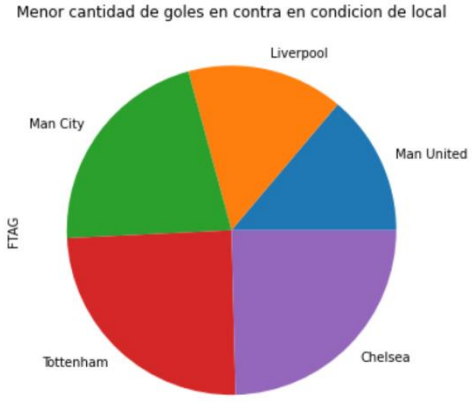
Yaya Toure	Manchester+City	86	CM
Fernandinho	Manchester+City	78	DM
Ilkay Gundogan	Manchester+City	41	CM

Defensa y portería

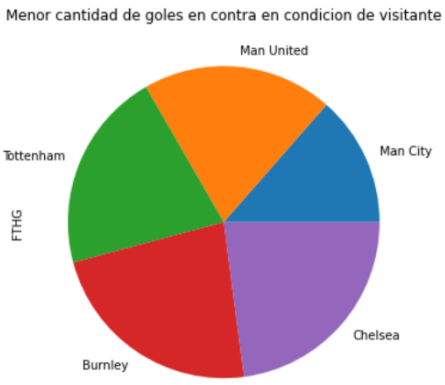
Los goles en contra y la cantidad de faltas cometidas indican que tan capaz es la defensa y el portero de defender el equipo sin la necesidad de cometer faltas.



Por medio de esta grafica se puede observar los 5 equipos con más “Clean sheets matches” que traduce los partidos donde el club no han recibido ningún gol.



Estos gráficos indican los 5 equipos con menor cantidad de goles en contra tanto de visitante como de local. En ellas equipos como el Man. City, Tottenham y Chelsea cuentan con la menor cantidad de goles en contra



Listado de posibles jugadores de los defensas con menor cantidad de goles en contra

Kyle Walker	Manchester+City	142	RB
Nicolas Otamendi	Manchester+City	100	CB
Aleksandar Kolarov	Manchester+City	95	LB
John Stones	Manchester+City	59	CB
Vincent Kompany	Manchester+City	57	CB

Gary Cahill	Chelsea	178	CB
Marcos Alonso Mendoza	Chelsea	177	LB
Cesar Azpilicueta	Chelsea	170	RB
David Luiz	Chelsea	132	CB
Kurt Zouma	Chelsea	15	CB

Jan Vertonghen	Tottenham	126	CB
Toby Alderweireld	Tottenham	120	CB
Ben Davies	Tottenham	90	LB
Danny Rose	Tottenham	84	LB
Kieran Trippier	Tottenham	51	RB

Listado de posibles porteros teniendo en cuenta los últimos 3 gráficos

Simon Mignolet	Liverpool	110	GK

Claudio Bravo	Manchester+City	73	GK

Thibaut Courtois	Chelsea	141	GK

Conclusiones y Hallazgos

A lo largo del análisis realizado fue posible observar ciertas tendencias que fueron poco a poco se hicieron más repetitivas en los equipos que obtuvieron una alta puntuación para así quedar arriba en la tabla de posiciones de la Premier league en el año 2017 -18.

- Como estilo de juego se recomienda intentar mantener la posesión de la pelota pues equipos como el Manchester City, Arsenal y Chelsea que contaron con una posesión mayor al 60% del total de los partidos acumulada obtuvieron a la vez una menor cantidad de goles en contra
- El equipo con el portero que tuvo mayor número de partidos sin ningún gol no esta dentro de los 4 equipos en recibir menos goles en total en condición tanto de visitante como de local. Esto puede indicar que el portero es habilidoso y los errores son por parte de la defensa.
- Las delanteras mas solidas en cuanto a llegadas y efectividad son las del Arsenal y el Manchester City con
- Hay relación entre la posesión de la pelota y la posición en la tabla al final de la liga, los 5 equipos con más posesión de pelota están dentro de los 6 primeros puestos del resultado final de la liga
- Equipos como el Tottenham posee una gran diferencia en los goles recibidos cuando juega de local a que cuando juega de visitante donde de local es el equipo que recibe menos goles de todos.

Enlace para ver el resultado de la liga completo: https://www.fichajes.com/inglaterra/premier-league/2017-2018/clasificacion