Procesamiento de datos

Miguel Méndez Hernández

Parcial 1

Informe ejecutivo

Neyl Peñuela y Victoria Chavarro

Septiembre 18 de 2023

Contexto general del problema

Hemos sido contactados por un inversor millonario con el objetivo de crear un nuevo equipo de fútbol en la liga inglesa. El inversor tiene un gran interés en comprender en detalle la liga 17-18, incluyendo los equipos participantes, los resultados de los partidos y las estadísticas relacionadas con los equipos y jugadores. Esta información es crucial para que el inversor pueda tomar decisiones informadas sobre la dirección que debe tomar el nuevo equipo.

El inversor desea obtener una visión integral de la liga 17-18 inglesa, lo que incluye:

1. *Equipos Participantes:* Es necesario recopilar una lista de todos los equipos que compitieron en la liga. Esto incluye detalles sobre la ubicación geográfica de los equipos y su historial en temporadas anteriores.
2. *Resultados de los Partidos:* Se requiere un registro completo de los resultados de todos los partidos de la temporada. Esto incluye información sobre los equipos que se enfrentaron, la fecha y el lugar de los partidos, así como el resultado final (victoria, empate o derrota) de cada equipo.
3. *Estadísticas de los Equipos:* Es esencial recopilar estadísticas detalladas sobre el rendimiento de cada equipo durante la temporada. Esto podría incluir datos como la cantidad de goles anotados y encajados, la posición en la tabla de clasificación, la cantidad de victorias, empates y derrotas, entre otros indicadores clave.
4. *Estadísticas de los Jugadores:* Además de las estadísticas de los equipos, también es importante obtener información detallada sobre el rendimiento de los jugadores destacados de la temporada. Esto podría abarcar datos como la cantidad de goles anotados, asistencias, tarjetas amarillas y rojas, entre otros.

El inversor utilizará esta información para tomar decisiones estratégicas sobre la creación del nuevo equipo de fútbol. Esto incluye la identificación de jugadores talentosos que puedan ser contratados, el estilo de juego que el equipo debería adoptar y la formulación de una estrategia para competir en la liga inglesa.

Nuestra tarea principal es recopilar, analizar y presentar de manera clara y concisa toda la información relevante sobre la liga 17-18 inglesa para ayudar al inversor a tomar decisiones informadas y exitosas en la creación de su nuevo equipo de fútbol. Este informe proporcionará una visión detallada de la temporada y servirá como base para la toma de decisiones estratégicas.

Análisis de los conjuntos de datos y sus variables más importantes

Nombre del Conjunto de Datos: Players

- *Descripción:* Este conjunto de datos contiene información sobre jugadores de la Premier League, incluyendo detalles como su nombre, club, edad, posición, valor de mercado, puntos en el juego FPL, y más. Los datos están organizados en filas, donde cada fila representa a un jugador y las columnas contienen información específica sobre ese jugador. El conjunto de datos es útil para el análisis y la toma de decisiones relacionadas con el rendimiento de los jugadores en la liga.
- *Variables importantes seleccionadas y sus descriptivas generales*: Para seleccionar las 6 variables más importantes de este conjunto de datos con el objetivo de crear un nuevo equipo de fútbol en la Liga Inglesa, es importante considerar las características que son relevantes para tomar decisiones estratégicas sobre jugadores y estilo de juego. A continuación, se presentan las 6 variables seleccionadas:

1. Club: La pertenencia de un jugador a un club es crucial para formar un equipo. El club puede influir en la disponibilidad de un jugador y su estilo de juego.
2. Age (Edad): La edad de un jugador es importante para evaluar su potencial de desarrollo y su capacidad actual. Jugadores más jóvenes pueden tener un mayor potencial de crecimiento.
3. Position (Posición): La posición en la que juega un jugador es fundamental para construir un equipo equilibrado. Diferentes posiciones tienen diferentes roles y habilidades.
4. Market Value (Valor de mercado): El valor de mercado de un jugador puede indicar su calidad y demanda en el mercado de fichajes. Jugadores con un alto valor de mercado a menudo tienen un desempeño destacado.
5. FPL Points (Puntos de Fantasy Premier League): Estos puntos reflejan el desempeño de un jugador en la Fantasy Premier League y pueden indicar su capacidad para acumular puntos en el juego, lo que podría ser relevante para un equipo.
6. Name (nombre): Para poder identificar si el nombre del jugador es conocido por su desempeño.

Estas variables proporcionan una base sólida para tomar decisiones sobre la creación de un nuevo equipo de fútbol en la Liga Inglesa, teniendo en cuenta la calidad de los jugadores, su edad, sus posiciones, su valor de mercado, su historial de puntos en Fantasy Premier League y su nombre.

- *Reporte de calidad del conjunto de datos:*

Datos Faltantes: Se observa que no hay datos faltantes en ninguna de las columnas, lo que es positivo para la calidad de los datos. *Precisión de Datos:* La columna "age" (edad) parece estar completa y precisa, ya que las edades están dentro de un rango razonable para jugadores de fútbol. La columna "market_value" (valor de mercado) contiene valores numéricos que parecen ser precisos y coherentes con el contexto. *Integridad de Datos:* Las columnas "name" (nombre) y "nationality" (nacionalidad) no tienen

valores nulos y contienen información completa y coherente. *Consistencia de Datos:* Las columnas "position" (posición) y "position_cat" (categoría de posición) parecen estar en línea y consistentes con respecto a las posiciones de los jugadores. La columna "fpl_sel" (selección del juego Fantasy Premier League) se presenta como porcentaje en formato de cadena, lo que puede dificultar su uso en cálculos numéricos. Sería recomendable convertirla a un formato decimal para facilitar su análisis. *Valores Atípicos:* No se observan valores atípicos evidentes en ninguna de las columnas numéricas, como "age" o "market_value". *Duplicados:* No se identifican registros duplicados en el conjunto de datos. *Recomendaciones:* Convertir la columna "fpl_sel" a un formato decimal para facilitar el análisis. Realizar un análisis más detallado de las columnas categóricas, como "position" y "nationality", para identificar tendencias o patrones interesantes. Y Explorar la relación entre el "valor de mercado" y otras variables, como la "edad" o la "posición", para comprender mejor los factores que influyen en el valor de mercado de los jugadores.

Sample Data

	name	club	age	position	position_cat	market_value	page_views	fpl_value	fpl_sel	fpl_points	region	nationality	new_foreig
1	Alexis Sanchez	Arsenal	28	LW	1	65	4329	12	17.10%	264	3	Chile	0
2	Mesut Ozil	Arsenal	28	AM	1	50	4395	9.5	5.60%	167	2	Germany	0
3	Petr Cech	Arsenal	35	GK	4	7	1529	5.5	5.90%	134	2	Czech Republic	0
4	Theo Walcott	Arsenal	28	RW	1	20	2393	7.5	1.50%	122	1	England	0
5	Laurent Koscielny	Arsenal	31	CB	3	22	912	6	0.70%	121	2	France	0
6	Hector Bellerin	Arsenal	22	RB	3	30	1675	6	13.70%	119	2	Spain	0
7	Olivier Giroud	Arsenal	30	CF	1	22	2230	8.5	2.50%	116	2	France	0

Schema

	col_name	data_type	comment
1	name	string	null
2	club	string	null
3	age	int	null
4	position	string	null
5	position_cat	string	null
6	market_value	int	null
7	page_views	int	null

Nombre del Conjunto de Datos: Results

- *Descripción:* El conjunto de datos proporciona información sobre partidos de fútbol de la temporada. Contiene diversas características relevantes para cada partido, incluyendo detalles sobre la fecha y hora del partido, los equipos locales y visitantes (HomeTeam y AwayTeam), el número de goles anotados por ambos equipos en el tiempo completo (FTHG y FTAG), el resultado final del partido (FTR), el número de goles anotados por ambos equipos en el medio tiempo (HTHG y HTAG), el resultado del medio tiempo (HTR), el nombre del árbitro, los tiros a puerta (HS y AS), los tiros a puerta en el medio tiempo (HST y AST), los corners (HC y AC), las faltas (HF y AF), las tarjetas amarillas (HY y AY) y las tarjetas rojas (HR y AR).
- *Variables importantes seleccionadas y sus descriptivas generales*: Para seleccionar las variables más importantes de este conjunto de datos en el contexto del inversor millonario que quiere crear un nuevo equipo de fútbol en la Liga Inglesa y necesita información relevante, consideraría las siguientes 6 variables:

1. Season: Esta variable es esencial para identificar la temporada específica de los datos. Permite al inversor conocer en qué año se jugaron los partidos y cómo ha evolucionado la liga a lo largo del tiempo.
2. HomeTeam y AwayTeam: Estas variables indican los equipos que compitieron en cada partido. Son fundamentales para identificar el desempeño de los equipos y evaluar su estilo de juego.
3. FTHG (Full-Time Home Goals) y FTAG (Full-Time Away Goals): Estas variables registran la cantidad de goles que anotó el equipo local (HomeTeam) y el equipo visitante (AwayTeam) en un partido. Son cruciales para evaluar el rendimiento ofensivo y defensivo de los equipos y determinar su capacidad para marcar y evitar goles.
4. FTR (Full-Time Result): Esta variable muestra el resultado final de un partido, ya sea una victoria local (H), una victoria visitante (A) o un empate (D). Es importante para comprender cómo se desempeñaron los equipos en términos de resultados.

Este conjunto de datos es valioso para realizar análisis y estudios relacionados con el fútbol, como la predicción de resultados, el análisis de rendimiento de equipos y jugadores, y la evaluación del desempeño de los árbitros. Cada fila representa un partido individual, lo que permite un análisis detallado de la temporada de la liga de fútbol en cuestión.

- *Reporte de calidad del conjunto de datos:*

Integridad de los datos: En general, los datos parecen estar completos, ya que no faltan valores en las columnas esenciales, como la fecha, los equipos, los goles y los resultados (FTHG, FTAG y FTR). *Precisión de los datos:* Las columnas de goles (FTHG y FTAG) parecen contener valores precisos que representan los goles marcados por los equipos en cada partido. Las columnas de resultados (FTR) también parecen precisas, ya que contienen las categorías adecuadas (H para victoria local, A para victoria visitante, D para empate). Las fechas y horas parecen estar en un formato correcto (columna DateTime). *Consistencia de los datos:* Las columnas de resultados (FTR) siguen un patrón consistente con categorías específicas (H, A, D), lo que facilita su análisis. Los nombres de los equipos y los árbitros no muestran problemas de consistencia evidentes. *Confiabilidad de los datos:* La confiabilidad de los datos no se puede evaluar completamente sin información adicional sobre la fuente y el proceso de recopilación de datos. Es importante verificar la fuente de los datos para determinar su confiabilidad. *Datos faltantes:* Algunas columnas, como HTHG, HTAG y HTR, contienen "NA" (No disponible) en todos los registros. Esto indica que los datos no están presentes en el conjunto proporcionado. *Limpieza de datos:* Es posible que sea necesario realizar una limpieza adicional de datos para eliminar o tratar los valores "NA" si se planea utilizar estas columnas en análisis posteriores.

	Sample Data	Schema

Sample Data:

	Season	DateTime	HomeTeam	AwayTeam	FTHG	FTAG	FTR	HTHG	HTAG	HTR	Referee	HS	AS
1	1993-94	1993-08-14T00:00:00.000+0000	Arsenal	Coventry	0	3	A	null	null	NA	NA	null	null
2	1993-94	1993-08-14T00:00:00.000+0000	Aston Villa	QPR	4	1	H	null	null	NA	NA	null	null
3	1993-94	1993-08-14T00:00:00.000+0000	Chelsea	Blackburn	1	2	A	null	null	NA	NA	null	null
4	1993-94	1993-08-14T00:00:00.000+0000	Liverpool	Sheffield Weds	2	0	H	null	null	NA	NA	null	null
5	1993-94	1993-08-14T00:00:00.000+0000	Man City	Leeds	1	1	D	null	null	NA	NA	null	null
6	1993-94	1993-08-14T00:00:00.000+0000	Newcastle	Tottenham	0	1	A	null	null	NA	NA	null	null
7	1993-94	1993-08-14T00:00:00.000+0000	Oldham	Ipswich	0	3	A	null	null	NA	NA	null	null

	col_name	data_type	comment
1	Season	string	null
2	DateTime	timestamp	null
3	HomeTeam	string	null
4	AwayTeam	string	null
5	FTHG	int	null
6	FTAG	int	null
7	FTR	string	null

Nombre del Conjunto de Datos: Season stats

- *Descripción:* Esta base de datos contiene información específica de equipos de futbol que hacen parte de la liga inglesa, en los que se proporcionan detalles específicos sobre su rendimiento en los partidos jugados en la temporada. Siendo estos 380 partidos. Mostrando estadísticas de rendimiento individual de los jugadores en toda la temporada, tal como los pases exitosos realizados en los partidos que jugó.
- *Variables importantes seleccionadas y sus descriptivas generales*: Para ayudar al inversor millonario a tomar decisiones informadas sobre la creación de un nuevo equipo de fútbol en la Liga Inglesa, es importante seleccionar las variables más relevantes que proporcionen una comprensión sólida de los equipos y los resultados de esa temporada. Claro estos van a variar teniendo en cuenta la posición, y a su vez,

  Aquí están las 6 variables más importantes que deberíamos considerar:

1. possession_percentage: El porcentaje de posesión de balón es una estadística clave para comprender el estilo de juego de un equipo. Ayudará al inversor a determinar si un equipo prefiere un enfoque de posesión o un juego más directo.
2. total_tackle: La cantidad total de entradas o tackles realizados por un equipo indica su capacidad defensiva y su habilidad para recuperar el balón. Ayudará al inversor a evaluar la solidez defensiva de los equipos.
3. team_name: Esta variable es esencial para identificar los equipos y conocer sus nombres. Ayudará al inversor a conocer la historia, la reputación y el mercado potencial de cada equipo en la liga.
4. accurate_pass: La cantidad de pases precisos realizados por un equipo es crucial para evaluar su habilidad en la distribución del balón y la creación de oportunidades de gol.
5. total_scoring_att: El número total de intentos de anotación refleja la capacidad ofensiva de un equipo. Esto es esencial para identificar equipos que tienden a ser más ofensivos.
6. team_rating: La calificación del equipo es crucial para comprender el nivel de rendimiento esperado. Esto ayudará al inversor a identificar equipos exitosos y aspiraciones competitivas.

Estas variables proporcionarán al inversor una comprensión sólida de los equipos, sus estilos de juego, su rendimiento y sus resultados en la temporada de la Liga Inglesa. Con esta información, el inversor podrá tomar decisiones más informadas sobre qué tipo de jugadores preferiría contratar y qué estilo de juego le gustaría que adoptara su nuevo equipo.

- *Reporte de calidad del conjunto de datos:*

Integridad de los datos: Los datos parecen estar completos y no faltan valores críticos en los registros. *Precisión de los datos:* Las métricas numéricas, como "team_rating," "won_corners," "total_tackle," "possession_percentage," etc., parecen estar en el formato correcto y son numéricas, lo que indica precisión. Las fechas parecen estar en el formato correcto ("dd/mm/yyyy"). *Consistencia de los datos:* Los datos siguen una estructura jerárquica consistente con información detallada sobre equipos y jugadores. Los nombres de los equipos y jugadores están formateados de manera uniforme. Las métricas parecen estar en la unidad correcta y siguen un patrón constante. *Homogeneidad:* Los datos siguen un formato homogéneo en toda la estructura del archivo JSON.

Este archivo se subdivide en 2 partes, una en la que se toman solo las estadísticas del equipo en general, y la otra en las estadísticas de los jugadores.

Jugador:

	Sample Data	Schema

	|	player_id	player_name	player_position_value	player_position_info	player_rating	own_goals	good_high_claim	penalty_conceded	touches	error_lead
1	Jonas Lössl	131171	Jonas Lössl	1	GK	6.88	null	null	null	33	null
2	null	131171	Jonas Lössl	1	GK	5.75	null	null	null	41	null
3	null	131171	Jonas Lössl	1	GK	5.75	null	null	null	41	null
4	null	131171	Jonas Lössl	1	GK	7.28	null	null	null	31	null
5	null	131171	Jonas Lössl	1	GK	8.1	null	null	null	40	null
6	Terence Kongolo	109117	Terence Kongolo	5	Sub	6.78	null	null	null	28	null
7	null	109117	Terence Kongolo	2	DL	7.16	null	null	null	81	null

	col_name	data_type	comment
1	|	string	null
2	player_id	string	null
3	player_name	string	null
4	player_position_value	int	null
5	player_position_info	string	null
6	player_rating	float	null
7	own_goals	int	null

Equipo:

	Sample Data	Schema

	|	team_id	team_name	team_rating	date	won_corners	att_sv_low_centre	won_contest	att_goal_high_right	total_tackl
1	166	166	Huddersfield	6.8828573	null	4	1	11	null	27
2	null	166	Huddersfield	6.1814284	null	1	null	8	null	8
3	null	166	Huddersfield	6.8557143	null	1	1	3	null	18
4	null	166	Huddersfield	6.4085712	null	2	null	4	null	22
5	26	26	Liverpool	6.4135714	null	13	1	16	null	13
6	null	26	Liverpool	7.252857	null	10	4	15	null	11
7	null	26	Liverpool	7.495714	null	7	3	13	null	26

	col_name	data_type	comment
1	|	string	null
2	team_id	string	null
3	team_name	string	null
4	team_rating	float	null
5	date	timestamp	null
6	won_corners	int	null
7	att_sv_low_centre	int	null

Visualizaciones descriptivas de las variables y hallazgos interesantes del negocio

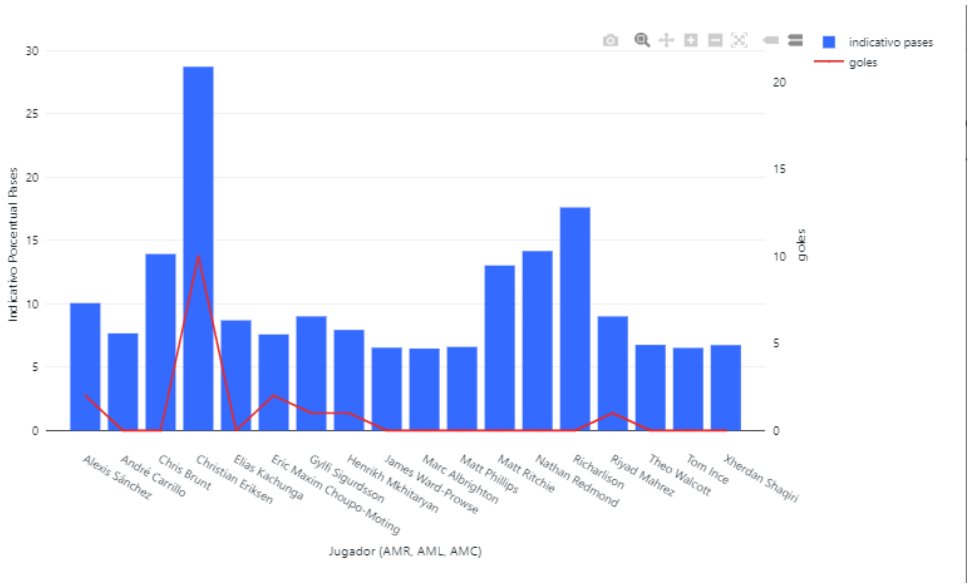
Para encontrar los prospectos del fútbol inglés mejor equipo de fútbol con los jugadores ya participantes de la liga inglesa, se decidió tomar en cada posición diferentes estadísticas, permitiendo así jerarquizar cuáles son los mejores jugadores para cada posición.

Para el estudio y filtro de las siguientes posiciones se presentó la necesidad de cuantificar el peso participativo de pases certeros en todo el equipo, esto ya que su posición al encontrarse en el medio es de dónde se crean las oportunidades de gol y en sí, la mayor aptitud o limitante que permite a los equipos ganar partidos.

Para lograr lo anterior explicado, se realizó un indicativo de "participación precisa de pases en el equipo" de cada jugador, donde se mide de manera porcentual su asertividad en sus pases sobre todos sus intentos de pases, con respecto a la cantidad de pases generadas en todos los juegos de la temporada; Siendo su finalidad encontrar esos jugadores que pueden ser la base de un equipo:

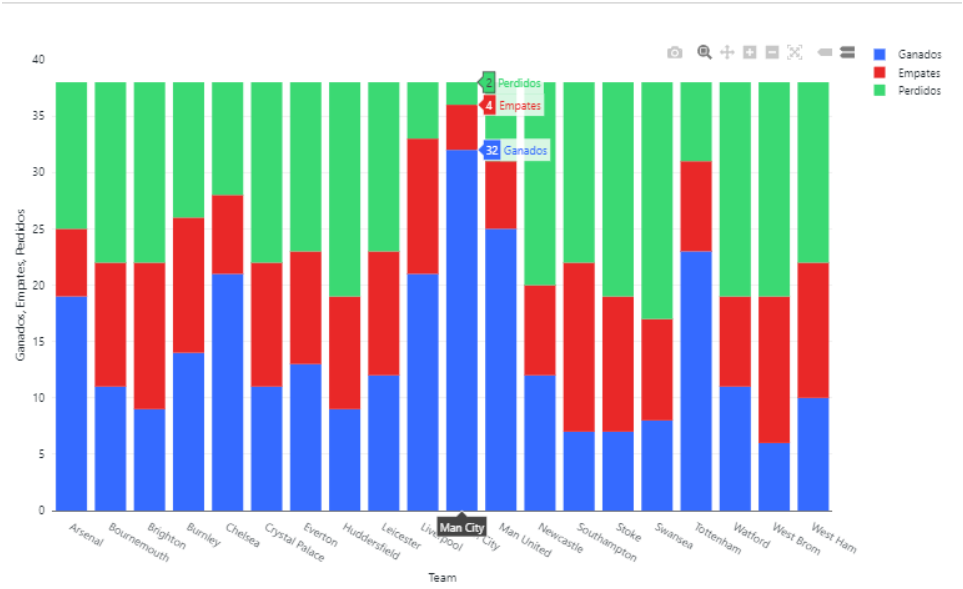
```
((JUga.pases_j - JUga.acc_j) * 100) / (EQu.pases_e - EQu.acc_e) AS IND
```

Tal como muestra el siguiente ejemplo, que analiza la cantidad de asistencias que tuvieron ciertos jugadores al lado de su respectivo valor del indicativo ya explicado:



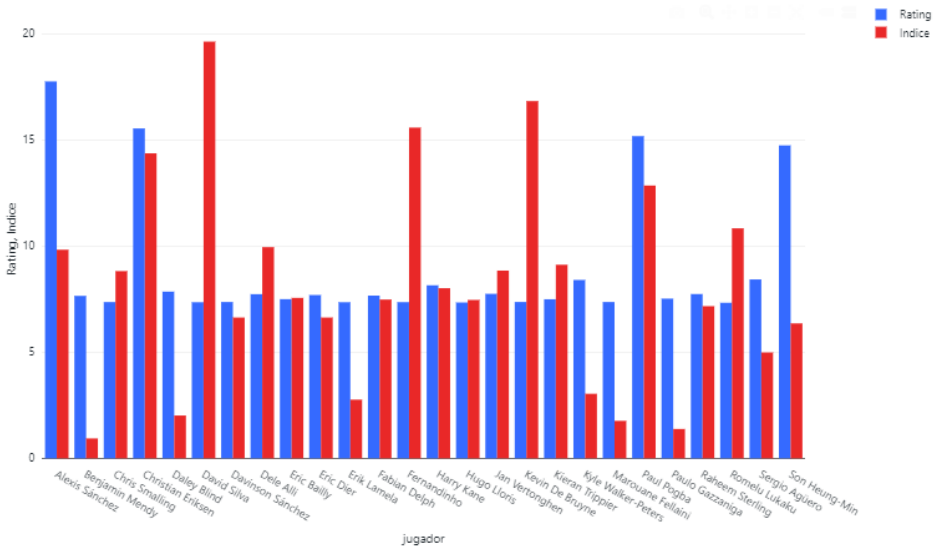
Aquí se muestra de los mejores 18 Mediocampistas de ataque (AML, AMR, AMC)

Al hacer el estudio con el archivo results de todos los partidos jugados en la Temporada 2017-2018 nos encontramos con esto:



Manchester City fue el equipo que más partidos ganó en toda la temporada, siguiendo de cerca el Manchester United y luego el TottenHam. Con esto en mente, se decidió encontrar los mejores

jugadores entre estos equipos al medir su promedio de la calificación y ver su relación con el indicativo hecho.



Al observar el anterior gráfico de barras, se encuentran puntos donde se asemejan los niveles que muestra el índice y el “rating”. Permitiéndonos observar la relación que hay entre los pases y la calificación, esta, tornándose como un punto determinante para entender si tener un buen control de pases es un indicador de una buena calificación (“rating”). Esto, da a entender que para esta calificación tener un control de pases no es lo más importante, puesto que los resultados, como asistir, tapar y el marcar goles lo llegan a ser, no solo un buen rendimiento; es así, el por qué están los goleadores en el foco de esa grafica.

Por otro lado, esto demostraría que tal vez no tomándose en cuenta en este “rating” que algunos pases dan apertura a asistencias, y estos a goles; lo cual demuestra que este índice toma como factor principal la creación de oportunidades de gol, como un potenciador de estos resultados, y por ende, una estadística necesaria. Esto puede ser afianzado al revisar los picos rojos (que corresponden al índice), encontrando que son jugadores cuyas calificaciones del momento no son las mejores, pero que hoy en día son de los jugadores que brillaron en la premier league en las temporadas siguientes. Esto invita a entender que hay jugadores que pasan por desapercibidos por una calificación y se tiende a ignorar su valor.

Conclusión

En este proyecto, hemos realizado un análisis exhaustivo de las variables clave en el fútbol: arquero, centrocampista, delantero y defensa, utilizando datos de la temporada 2017-2018 de la liga inglesa. Nuestro objetivo era proporcionar al inversor millonario la información necesaria para tomar decisiones informadas sobre la creación de un nuevo equipo de fútbol y definir su estilo de juego. A través de visualizaciones y análisis de estadísticas, hemos identificado las fortalezas y debilidades de los jugadores y equipos en cada una de estas posiciones. Por ejemplo, hemos destacado la importancia de la precisión en los pases para los centrocampistas, la capacidad de marcar goles para los delanteros y la eficacia en la defensa para los defensores. También hemos considerado el rendimiento de los arqueros en términos de evitar goles y realizar paradas cruciales.

Estos hallazgos proporcionan una base sólida para la toma de decisiones estratégicas. El inversor millonario ahora tiene una comprensión más clara de qué tipo de jugadores preferiría contratar para su equipo y cómo debería ser el estilo de juego del equipo para alcanzar el éxito en la liga inglesa. Además, este análisis puede servir como referencia para la planificación a largo plazo y la evaluación continua del rendimiento del equipo. En resumen, este proyecto ha proporcionado una valiosa visión de los datos de la temporada 2017-2018 de la liga inglesa, permitiendo al inversor millonario tomar decisiones informadas y estratégicas para la creación de su nuevo equipo de fútbol. El conocimiento obtenido aquí servirá como un activo invaluable en el proceso de establecer un equipo exitoso en el competitivo mundo del fútbol.