ANALISIS DE DATOS EN EL FUTBOL

Descubriendo el Rendimiento de Jugadores y Equipos

Autor: Luis Solis

Agenda

- 01 | Introducción
- 02 | Conjunto de Datos
- 03 | Preguntas e Hipótesis
- 04 | Objetivo – Contexto Comercial
- 05 | Análisis Exploratorio
- 06 | Respondiendo Preguntas
- 07 | Implementación de Modelos de Machine Learning
- 08 | Conclusiones

1. Introducción

En el mundo del fútbol, el análisis de datos ha cobrado una gran relevancia en los últimos años. La información obtenida a través del análisis de diversas variables puede proporcionar insights valiosos sobre el desempeño de los jugadores y equipos.

En este contexto, el presente conjunto de datos ofrece una amplia variedad de información que captura diferentes aspectos del juego.

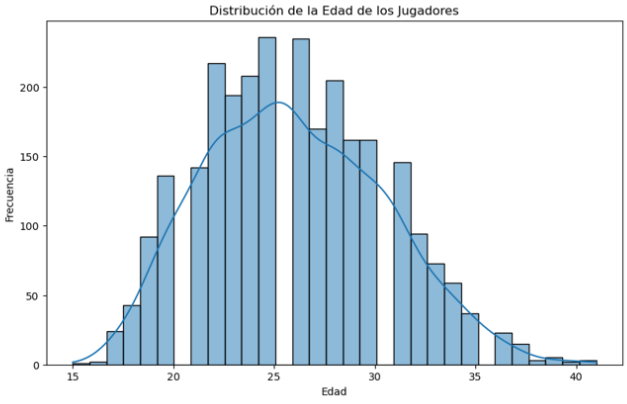
En este análisis, nos dirigimos a entrenadores, analistas y aficionados del fútbol interesados en comprender y utilizar estos datos para evaluar el rendimiento de los jugadores y equipos en diversos aspectos del juego.

2. Conjunto de Datos

El conjunto de datos que estamos analizando contiene una amplia variedad de variables que capturan diferentes aspectos del juego de fútbol. Estas variables incluyen goles marcados, precisión de pases, regates exitosos, habilidades defensivas, edad, posición en el campo, altura y más. Esta información es esencial para evaluar el rendimiento de los jugadores y equipos en diferentes aspectos del juego.

link Set de Datos: https://www.kaggle.com/datasets/ameyaranade/big-5-european-leagues-player-stats-2022-23

Resumen DataSet:



3. Preguntas e Hipótesis

Antes de sumergirnos en el análisis, estableceremos algunas preguntas e hipótesis que podríamos responder a través de los datos:

1. ¿Cuáles son los jugadores más efectivos en términos de goles marcados? ¿Existen diferencias significativas entre los distintos equipos en cuanto a la producción de goles?

2. ¿Existe una relación entre la precisión de los pases de un jugador y su capacidad para crear oportunidades de gol para su equipo?
3. ¿Cuál es el porcentaje de éxito de los regates realizados por los jugadores? ¿Algunos jugadores destacan por su habilidad para eludir a los defensores?

Estas preguntas guiarán nuestro análisis y nos ayudarán a obtener información relevante sobre el rendimiento de jugadores y equipos en el fútbol.

4. Objetivo – Contexto Comercial

4.a Objetivo

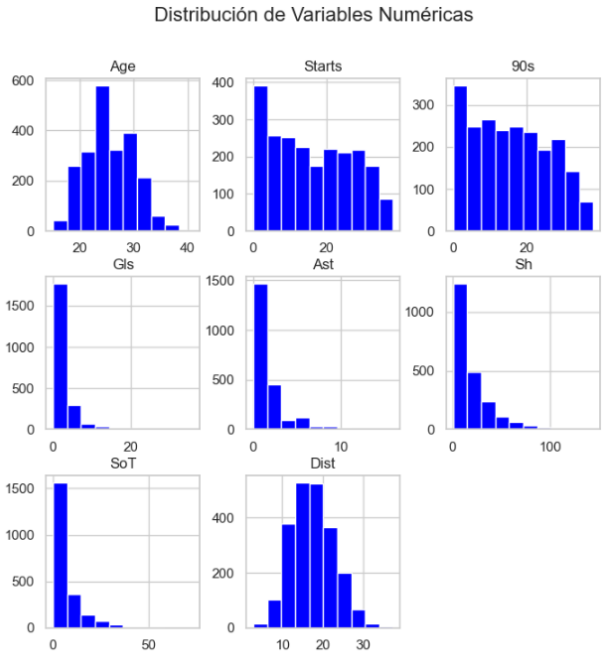
El objetivo de este análisis es utilizar técnicas de análisis de datos para examinar el rendimiento de jugadores y equipos de fútbol en función de diversas variables. Estas variables incluyen goles marcados, precisión de pases, regates exitosos, habilidades defensivas, edad, posición en el campo, altura y otras relevantes. Buscaremos identificar patrones y relaciones significativas que puedan proporcionar insights valiosos para la toma de decisiones estratégicas y operativas en la industria del fútbol.

4.b Contexto Comercial y Problema a Abordar

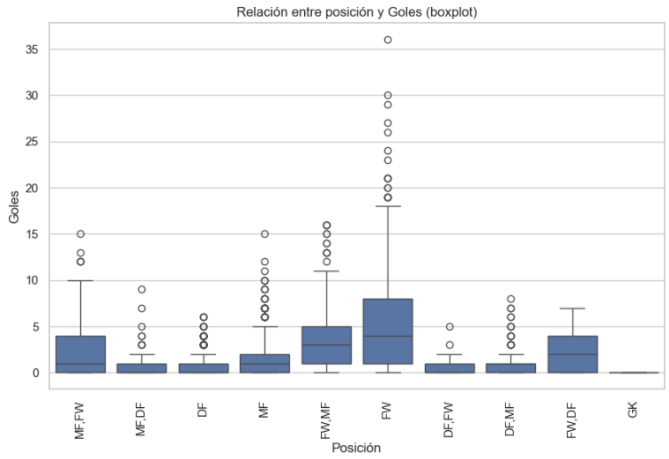
En el contexto comercial, nuestro objetivo es proporcionar información y análisis útiles para actores dentro de la industria del fútbol, como clubes, directores técnicos y agentes de jugadores. Abordamos el problema de comprender y evaluar el rendimiento de los jugadores y equipos de fútbol para facilitar la identificación de talentos, la selección de jugadores y la toma de decisiones estratégicas.

5. Análisis EDA

- Análisis univariado

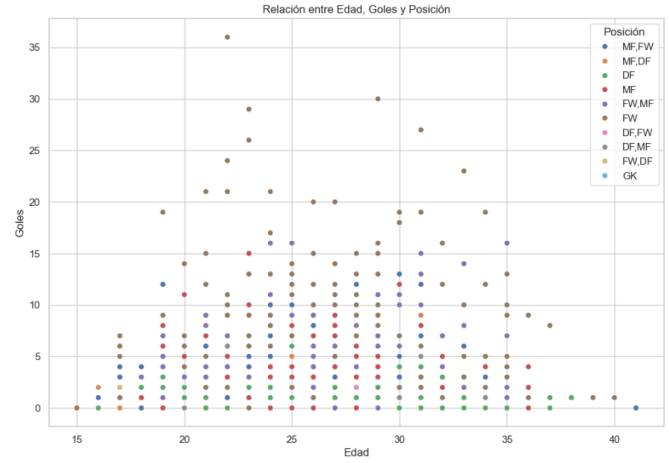


- Análisis Bivariado



En el grafico anterior, la posición FW es la que proporciona mayor cantidad de goles,lo cual es logico ya que es la posición mas cercana al arco contrario.

- Analisis trivariado



En el grafico anterior no se observa correlación entre edad y cantidad de goles. Si un rango minimo (20 años) y maximo (36 años) de promedio de goles.

Se observa ademas una correlación entre posición y cantidad de goles, la posición FW es la que proporciona mayor cantidad de goles,lo cual es logico ya que es la posición mas cercana al arco contrario.

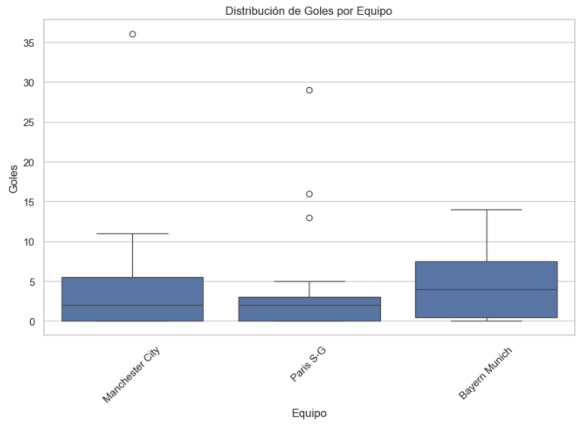
6. Respondiendo Preguntas

- ¿Cuáles son los jugadores más efectivos en términos de goles marcados?

	Player	Squad	Gls
578	Erling Haaland	Manchester City	36
762	Harry Kane	Tottenham	30
1159	Kylian Mbappé	Paris S-G	29
89	Alexandre Lacazette	Lyon	27
2080	Victor Osimhen	Napoli	26
980	Jonathan David	Lille	24
1778	Robert Lewandowski	Barcelona	23
657	Folarin Balogun	Reims	21
1173	Lautaro Martínez	Inter	21
1224	Loïs Openda	Lens	21
747	Habib Diallo	Strasbourg	20
838	Ivan Toney	Brentford	20
541	Elye Wahi	Montpellier	19
1083	Karim Benzema	Real Madrid	19
1490	Mohamed Salah	Liverpool	19

- ¿Existen diferencias significativas entre los distintos equipos en cuanto a la producción de goles?

```
            Squad  Gls
55  Manchester City   92
11    Bayern Munich   88
3           Arsenal   84
25         Dortmund   80
68        Paris S-G   79
63           Napoli   73
71      Real Madrid   73
49        Liverpool   70
10        Barcelona   68
39            Inter   68
```




```
Los 5 goleadores máximos del Bayern Munich:
           Squad                    Player  Gls
17  Bayern Munich              Serge Gnabry   14
6   Bayern Munich             Jamal Musiala   12
5   Bayern Munich  Eric Maxim Choupo-Moting   10
8   Bayern Munich            Kingsley Coman    8
10  Bayern Munich                Leroy Sané    8

Los 5 goleadores máximos del Paris S-G:
       Squad            Player  Gls
45  Paris S-G     Kylian Mbappé   29
46  Paris S-G      Lionel Messi   16
48  Paris S-G            Neymar   13
39  Paris S-G     Achraf Hakimi    5
41  Paris S-G  Fabián Ruiz Peña    3

Los 5 goleadores máximos del Manchester City:
             Squad           Player  Gls
22  Manchester City   Erling Haaland   36
32  Manchester City       Phil Foden   11
26  Manchester City   Julián Álvarez    9
38  Manchester City   İlkay Gündoğan    8
28  Manchester City  Kevin De Bruyne    7
```

Conclusiones

el rendimiento goleador de los equipos Manchester City, Paris S-G y Bayer Munich varía en función de la dependencia de ciertos jugadores. Mientras que el Manchester City se apoya fuertemente en Erling Haaland para la anotación de goles, Paris S-G confía en Kylian Mbappé, Neymar y Lionel Messi. Por otro lado, el equipo Bayer Munich muestra una distribución más equitativa en cuanto a la producción de goles, con varios jugadores contribuyendo de manera similar en términos de goles anotados".

Esta hipótesis sugiere que la dependencia de un solo jugador para la producción de goles puede influir en el rendimiento goleador de un equipo. Mientras que el Manchester City depende en gran medida de Erling Haaland, Paris S-G cuenta con varios jugadores clave, y el Bayer Munich muestra una mayor distribución en la producción de goles entre sus jugadores.

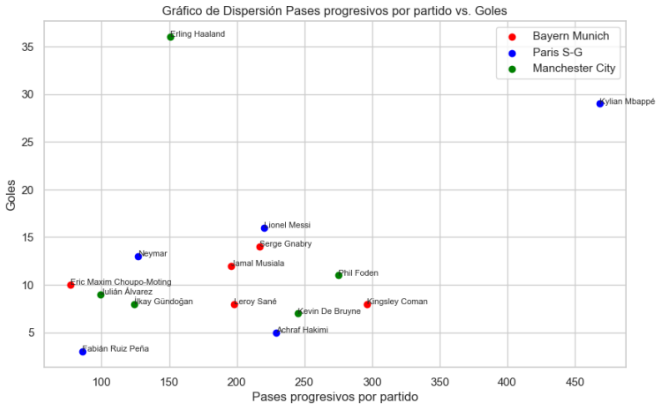
Si uno fuera rival de Manchester City, el planteo seria realizar una fuerte marcación contra Erling Haaland evitando que le llegue el balón.

- ¿Existe una relación entre la precisión de los pases de un jugador y su capacidad para crear oportunidades de gol para su equipo?

```
Los 5 goleadores máximos del Bayern Munich:
          Squad                  Player  Gls   PrgR
0  Bayern Munich            Serge Gnabry   14  217.0
1  Bayern Munich           Jamal Musiala   12  196.0
2  Bayern Munich  Eric Maxim Choupo-Moting   10   77.0
3  Bayern Munich          Kingsley Coman    8  296.0
4  Bayern Munich              Leroy Sané    8  198.0

Los 5 goleadores máximos del Paris S-G:
       Squad           Player  Gls   PrgR
0  Paris S-G     Kylian Mbappé   29  468.0
1  Paris S-G      Lionel Messi   16  220.0
2  Paris S-G           Neymar   13  127.0
3  Paris S-G     Achraf Hakimi    5  229.0
4  Paris S-G  Fabián Ruiz Peña    3   86.0

Los 5 goleadores máximos del Manchester City:
           Squad          Player  Gls   PrgR
0  Manchester City   Erling Haaland   36  151.0
1  Manchester City       Phil Foden   11  275.0
2  Manchester City   Julián Álvarez    9   99.0
3  Manchester City   İlkay Gündoğan    8  124.0
4  Manchester City  Kevin De Bruyne    7  245.0
```



Se calcula el coeficiente de correlación entre la precisión de los pases (PrgR) y la capacidad para crear oportunidades de gol (Gls). El resultado es 0.65.

Conclusiones

el coeficiente de correlación de 0.65 sugiere una correlación positiva moderada entre la precisión de los pases y la capacidad para crear oportunidades de gol. Sin embargo, es

necesario considerar otros factores y el contexto específico de cada jugador y equipo para obtener una comprensión más completa de su desempeño en el fútbol.

- ¿Cuál es el porcentaje de éxito de los regates realizados por los jugadores? ¿Algunos jugadores destacan por su habilidad para eludir a los defensores?

Para responder a las preguntas anteriores, utilizaría las siguientes variables:

ToAtt: Número de intentos de regatear a los defensores.

ToSuc: Número de defensores regateados con éxito.

ToSuc%: Porcentaje de regates completados con éxito.

	Player	Squad	ToAtt	ToSuc%
1233	Vinicius Júnior	Real Madrid	8.10	34.8
1348	Khvicha Kvaratskhelia	Napoli	6.86	28.7
2023	Nemanja Radonji?	Torino	6.67	29.4
1609	Kylian Mbappé	Paris S-G	6.63	31.5
500	Samuel Chukwueze	Villarreal	6.61	44.6
675	Stéphane Diarra	Lorient	6.48	33.9
1650	Lionel Messi	Paris S-G	6.29	53.6
2145	Georginio Rutter	Hoffenheim	6.16	46.4
1098	Luiz Henrique	Betis	5.97	40.3
256	Saïd Benrahma	West Ham	5.94	39.2

	Player	Squad	ToSuc	ToSuc%
1650	Lionel Messi	Paris S-G	3.37	53.6
2198	Leroy Sané	Bayern Munich	3.24	58.7
1866	Azzedine Ounahi	Angers	3.12	53.4
500	Samuel Chukwueze	Villarreal	2.95	44.6
1030	Evann Guessand	Nantes	2.95	52.5
2145	Georginio Rutter	Hoffenheim	2.86	46.4
1233	Vinicius Júnior	Real Madrid	2.82	34.8
606	Alphonso Davies	Bayern Munich	2.58	46.0
895	Jeremie Frimpong	Leverkusen	2.46	45.8
246	Jude Bellingham	Dortmund	2.45	49.5

Conclusiones

¿Cuál es el porcentaje de éxito de los regates realizados por los jugadores?

Los porcentajes de éxito de los regates se encuentran en la columna 'ToSuc%' en el DataFrame. Este valor representa la eficacia de los jugadores para eludir a los defensores en sus intentos de regate. Algunos de los mejores jugadores en términos de 'ToSuc%' son:

Lionel Messi (Paris S-G) con un 53.6% en 'ToSuc%'.

Leroy Sané (Bayern Munich) con un 58.7% en 'ToSuc%'.

Samuel Chukwueze (Villarreal) con un 44.6% en 'ToSuc%'.

Estos jugadores tienen un alto porcentaje de éxito en sus regates, lo que sugiere que son efectivos para eludir a los defensores en situaciones de uno contra uno.

¿Algunos jugadores destacan por su habilidad para eludir a los defensores?

Sí, algunos jugadores destacan por su habilidad para eludir a los defensores, lo que se refleja en su alto 'ToSuc%'. En particular, jugadores como Lionel Messi y Leroy Sané tienen un 'ToSuc%' destacado, lo que indica su habilidad para superar a los defensores con éxito en situaciones de regate. Estos jugadores son conocidos por su destreza en el dribbling y su capacidad para superar o eludir a los oponentes.

En resumen, los porcentajes de éxito de los regates ('ToSuc%') indican la eficacia de los jugadores para eludir a los defensores, y algunos jugadores sobresalen en esta habilidad, como

Lionel Messi y Leroy Sané. Estos jugadores son altamente efectivos en situaciones de regate y destacan en este aspecto del juego.

7. Implementación de Modelos de Machine Learning

Se realizan pruebas en los siguientes modelos:

- Regresion Lineal
- Random Forest
- Regresion Logistica
- Soporte Vectorial

Los resultados se describen a continuación:

El modelo de bosque aleatorio es el modelo más preciso para predecir los goles de un jugador de fútbol, con un RMSE de 0.318.

El modelo de regresión lineal es el modelo menos preciso, con un RMSE de 1.238.

El modelo de regresión logística es más preciso que el modelo de soporte vectorial, con una precisión de 0.632 frente a 0.496, respectivamente.

En general, los resultados obtenidos indican que los modelos de aprendizaje automático supervisado, como el bosque aleatorio y la regresión logística, son más precisos para predecir los goles de un jugador de fútbol que los modelos de aprendizaje automático no supervisado, como el soporte vectorial.

Los resultados también sugieren que los modelos que consideran la cantidad de disparos que realiza un jugador de fútbol y la precisión de esos disparos, así como la productividad goleadora en tiros directos y en tiros lejanos, son más precisos para predecir los goles de un jugador de fútbol.

Estos resultados pueden ser utilizados por los entrenadores de fútbol para identificar a los jugadores que tienen más probabilidades de marcar goles y para desarrollar estrategias para maximizar las posibilidades de gol de sus equipos.

- Analisis de Componentes Principales – PCA

Las cargas de los dos primeros componentes son las siguientes:

Componente 1: Está fuertemente correlacionado con las variables Sh, SoT, SoT_per_90 y Sh_per_90. Esto sugiere que este componente representa la cantidad de disparos que realiza un jugador de fútbol y la precisión de esos disparos.

Componente 2: Está fuertemente correlacionado con las variables Gls, FK y Dist. Esto sugiere que este componente representa la productividad goleadora de un jugador de fútbol, tanto en tiros directos como en tiros lejanos.

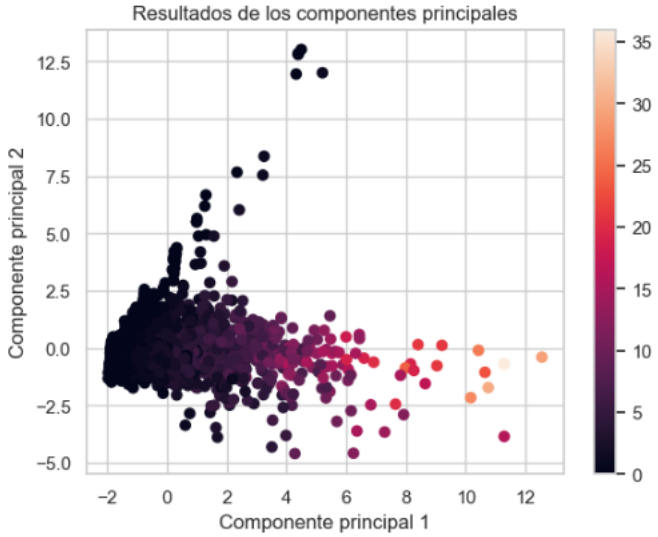
Las variables más relevantes para los dos primeros componentes son las siguientes:

Componente 1: Sh, SoT, SoT_per_90 y Sh_per_90

Componente 2: Gls, FK y Dist

En general, los resultados de la PCA indican que las variables más relevantes para predecir los goles de un jugador de fútbol son la cantidad de disparos que realiza, la precisión de esos disparos, la productividad goleadora en tiros directos y la productividad goleadora en tiros lejanos.

Además, los resultados de la PCA sugieren que las variables Age, Starts y 90s no están fuertemente correlacionadas con los goles marcados por un jugador de fútbol. Esto puede deberse a que estas variables representan factores que están fuera del control de un jugador, como la cantidad de tiempo que juega y la edad.



8. Conclusiones

- Metodo K-Fold Cross Validation para Modelo Random Forest Regressor

Las conclusiones para el método K-Fold Cross Validation son las siguientes:

El RMSE obtenido (0.4171) es mayor que el obtenido con el modelo original, lo que sugiere que el modelo es menos preciso cuando se utiliza K-Fold Cross Validation.

Este resultado puede deberse a que K-Fold Cross Validation utiliza una validación cruzada estratificada, lo que significa que cada partición tiene una distribución de clases similar a la distribución de clases en el conjunto de datos completo. En este caso, el conjunto de datos tiene dos clases: jugadores que marcaron goles y jugadores que no marcaron goles. Al utilizar K-Fold Cross Validation, nos aseguramos de que cada partición tenga un número similar de jugadores que marcaron goles y jugadores que no marcaron goles. Esto puede ayudar a mejorar la precisión del modelo de aprendizaje automático, pero también puede introducir sesgos en los resultados.

- Para mejorar la precisión del modelo con K-Fold Cross Validation, podemos considerar las siguientes soluciones:

Utilizar un número mayor de particiones. Esto ayudará a reducir el sesgo introducido por la validación cruzada estratificada.

Utilizar un método de validación no estratificado. Esto puede ayudar a mejorar la precisión del modelo en general, pero también puede introducir sesgos en los resultados.

Utilizar un modelo de aprendizaje automático diferente. Algunos modelos, como los modelos de regresión lineal, pueden ser más precisos que los modelos de regresión de bosque aleatorio cuando se utilizan métodos de validación cruzada.

- Metodo Stratified K-Fold para Modelo Random Forest Regressor

Las conclusiones para el método Stratified K-Fold son las siguientes:

El RMSE obtenido (0.3940) es mayor que el obtenido con el modelo Random Forest Regressor, lo que sugiere que el modelo es menos preciso.

Este resultado es esperado, ya que Stratified K-Fold toma en cuenta la distribución de las clases en los datos. En este caso, el conjunto de datos tiene dos clases: jugadores que marcaron goles y jugadores que no marcaron goles.

Al utilizar Stratified K-Fold, nos aseguramos de que cada partición tenga un número similar de jugadores que marcaron goles y jugadores que no marcaron goles. Esto puede ayudar a mejorar la precisión del modelo de aprendizaje automático, pero también puede reducir la precisión en los casos en que el modelo está bien entrenado para predecir una clase en particular.

Para mejorar la precisión del modelo utilizando Stratified K-Fold:

podemos aumentar el número de particiones. Esto ayudará a que el modelo tenga más datos de entrenamiento para cada clase.

Otra opción es utilizar un método de validación cruzada que no tome en cuenta la distribución de las clases. Por ejemplo, podemos utilizar K-Fold con un número de particiones impar. Esto ayudará a que el modelo tenga una partición completa para cada clase.