Análisis de Fuentes de Datos Diversas

Informe

Elaborado por Camilo Herradora

Descripción General de la fuente

Por medio de la plataforma sofifa.com, con un método llamado Web Scraping se logró recolectar un histórico detallado de los datos de jugadores de futbol a través del modo de juego "Modo Carrera" de la saga de videojuegos FIFA, desde la edición "FIFA 15" (correspondiente a datos de jugadores de la temporada 2014-15) hasta la edición "EA SPORTS FC 24" (correspondiente a datos de jugadores de la temporada 2023-24)

Sofifa es una plataforma que actúa como la base de datos de los jugadores, con cada estadística completa acerca del jugador y su desempeño en el videojuego.

A través del Web Scraping se almacenan:

- Todos los jugadores, entrenadores y equipos disponibles en FIFA 15, 16, 17, 18, 19, 20, 21, 22, 23 y también en EA Sports FC 24.
- Todas las actualizaciones de FIFA desde el 10 de septiembre de 2015 hasta el 22 de septiembre de 2023
- 109 atributos para jugadores, 8 atributos para entrenadores y 54 atributos para equipos
- URL de los jugadores, entrenadores y equipos eliminados
- Posiciones de jugadores, con el rol en el club y en la selección

- Atributos del jugador con estadísticas como ataque, habilidades, defensa, mentalidad, habilidades GK, etc.
- Datos personales del jugador como nacionalidad, club, fecha de nacimiento, salario, etc.
- Datos del equipo sobre sus entrenadores, su valor general y tácticas.

en los siguientes CSV, tipo de datos de dato semi-estructurado:

- female_coaches.csv
- female_players.csv
- female_teams.csv
- male_coaches.csv
- male_players.csv
- male_teams.csv

Dado que haremos un análisis orientado a los jugadores masculinos prospectos de las 4 entregas
"FIFA 21"
"FIFA 22"
"FIFA 23"
"EA FC 24"

utilizaremos
- male_players.csv

Resultados del análisis descriptivo con gráficos y tablas.

Se evaluaron numerosas variables y su significado.

El conteo de jugadores prometedores por edición:

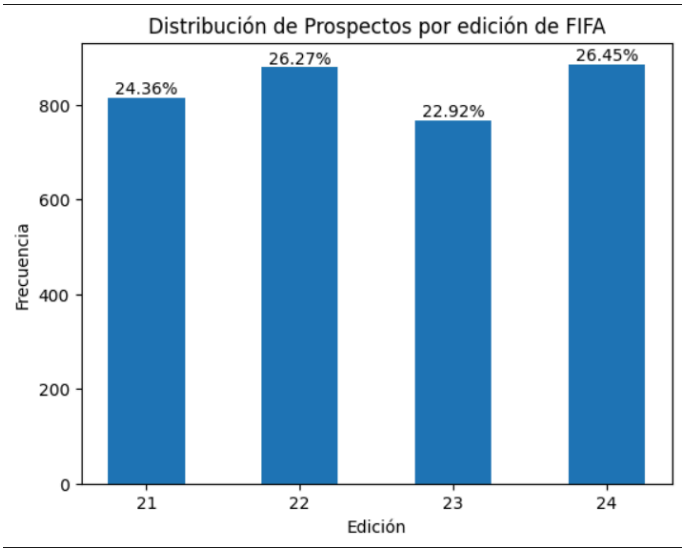
	fifa_version	count	percentage
0	24.0	885	26.449492
1	22.0	879	26.270173
2	21.0	815	24.357442
3	23.0	767	22.922893

Tanto en 2022 como 2024 fueron años en los que jóvenes promesas se destacaron para alcanzar un rango de “prospecto” en relación al 2021 y 2020.

```
#Creación de una nueva tabla, en la que creamos nuevos índices para los registros
edition_freq_table = prospects_df['fifa_version'].value_counts().reset_index()

#Columnas de la nueva tabla y columna nueva "Percentage"
edition_freq_table.columns = ['fifa_version', 'count']
edition_freq_table['percentage'] = (edition_freq_table['count'] / len(prospects_df)) * 100

edition_freq_table.head()
```



#Creación de gráfico de barras mostrando dicha distribución

```
plt.bar(edition_freq_table['fifa_version'],
edition_freq_table['count'], width = 0.5)

#Valores de la tabla de frecuencia para el eje x
plt.xticks(edition_freq_table['fifa_version'])

#Mostrando porcentajes por cada fila
for record, row in edition_freq_table.iterrows():
    plt.text(row['fifa_version'], row['count'],
f"{row['percentage']:.2f}%", ha='center', va='bottom')
```

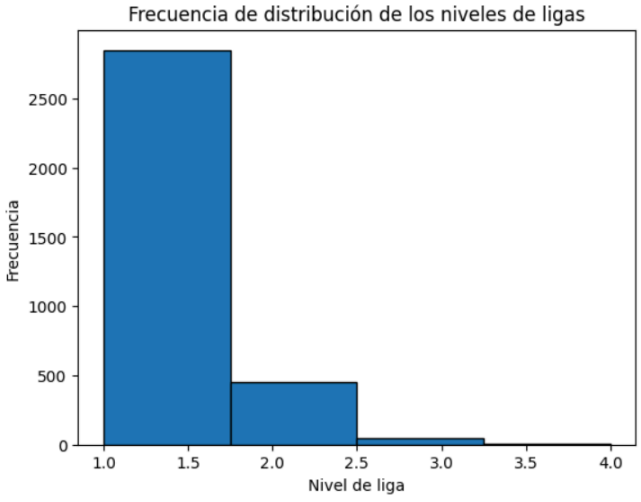
```
plt.title('Distribución de Prospectos por edición de FIFA')
plt.xlabel('Edición')
plt.ylabel('Frecuencia')
```

Se realizó un análisis descriptivo también de la variable “league_level”, cuya finalidad es describir en una escala ordinal (1 a 4) el nivel competitivo de una liga, de esta forma podríamos filtrar y lograr ver si existen jugadores jóvenes a un nivel competitivo de élite que provengan de las mejores ligas del mundo.

	league_level	count	percentage
0	1.0	2848	85.116557
1	2.0	451	13.478781
2	3.0	45	1.344889
3	4.0	2	0.059773

Con un 85%, se reflejaba que el mejor desempeño venía del grupo de jugadores al máximo rendimiento en las ligas más competitivas.

```
leagues_freq_table =
prospects_df['league_name'].value_counts().reset_index()
```

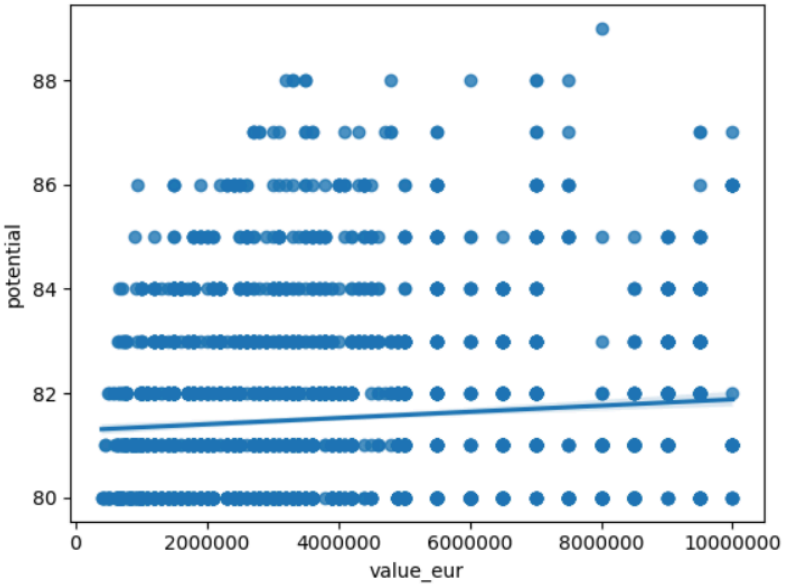



```
plt.hist(prospects_df['league_level'], bins=4,
edgecolor='black')
```

Discusión sobre patrones y relaciones encontradas.

El primer patrón identificado, es que a medida de que los jugadores prospectos crecen por la variable "potential", su valor en el mercado de transferencias crece, teniendo la frecuencia de valores para la tasa del mercado más apegada a aquellos jugadores con un valor de potencial decreciente.

El crecimiento del futbolista es directamente proporcional a su valor en el mercado de transferencias.



plt.gca().yaxis.set_major_formatter(ScalarFormatter(useOffset=False))
plt.ticklabel_format(style='plain', axis='x')

sns.regplot(x = 'value_eur', y = 'potential', data = prospects_df)

Otra correlación relevante y que merece la pena su mención, es aquella entre los potenciales, ,las ligas provenientes de cada jugador y el nivel de liga de estos últimos.

	league_name	count	percentage
0	Premier League	402	12.014345
1	Bundesliga	311	9.294680
2	La Liga	309	9.234907
3	Ligue 1	255	7.621040
4	Serie A	250	7.471608

En este caso, se ha filtrado por conteo de ligas que se repiten, cuáles son aquellas de las que más provienen los jugadores prospectos.

Mundialmente es sabido, que las mejores ligas del mundo, mejor conocidas como **"las 5 grandes ligas"** están en Europa y que son la cuna de muchos futbolistas exitosos actualmente.

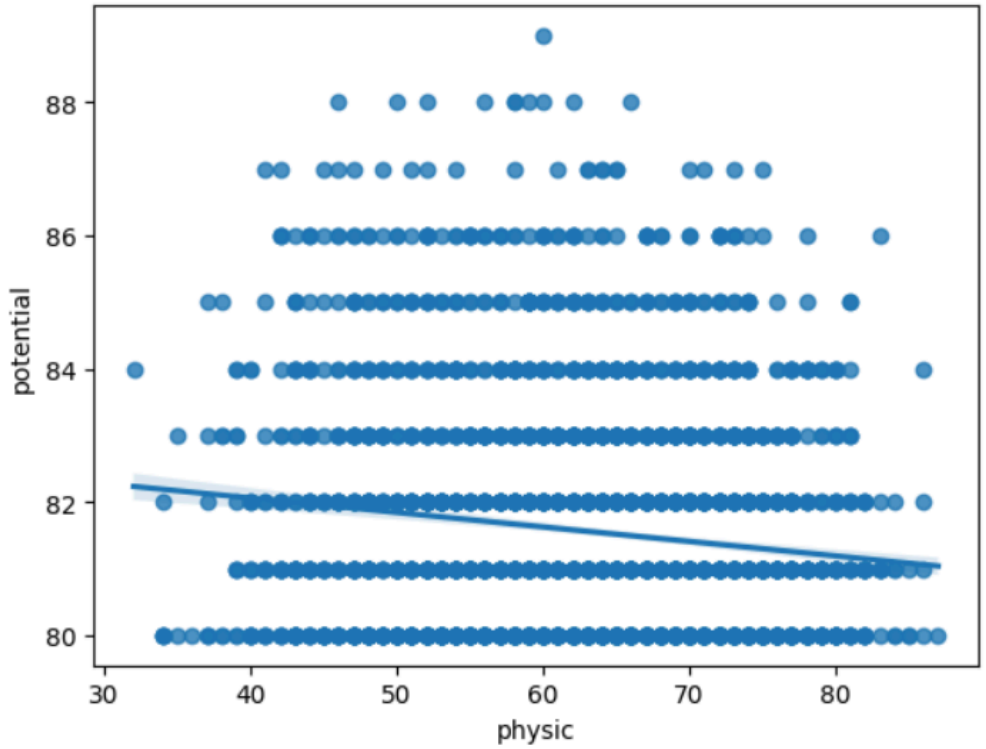
Las 5 grandes ligas son:
- Premier League
- LaLiga
- Bundesliga
- Ligue 1
- Serie A

Nuestra tabla de frecuencia de cuáles son las ligas con una mayor repetición en registros coincide exactamente con esta descripción anteriormente hecha.

¿Existen correlaciones entre variables?

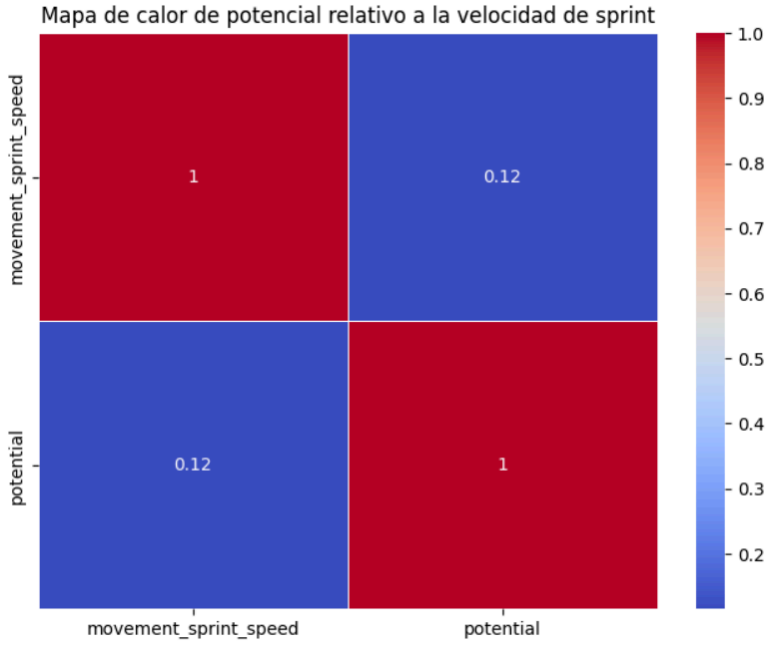
A parte de los factores ya estudiados y expuestos gráficamente (la liga, el nivel de liga, la relación entre el valor del jugador y el potencial que tiene), considero pertinente evaluar atributos físicos, qué tanta relación existe entre el potencial y rasgos como el físico, aceleración o velocidad.

Por otra parte veremos si existen patrones de comportamiento entre los valores de mercado y potencial



No hay tal correlación entre el valor de mercado y el potencial que marca el juego acerca del jugador.

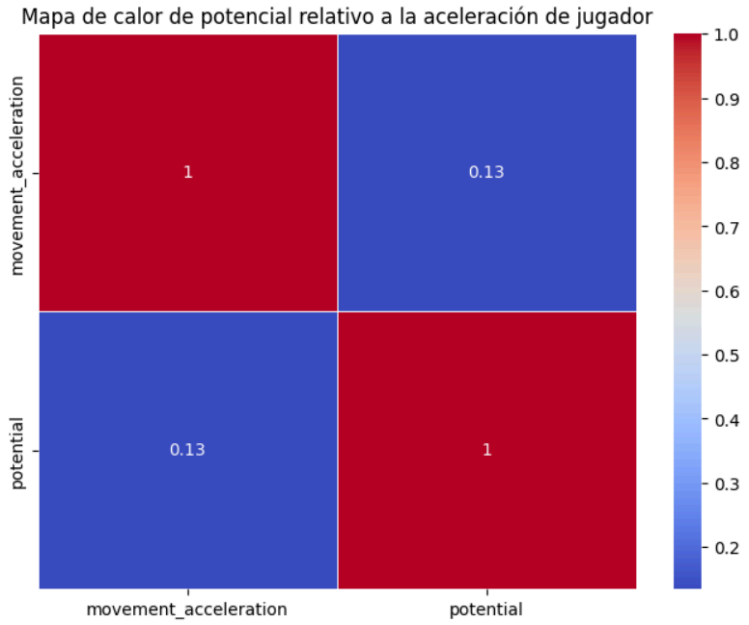
De acuerdo a las funcionalidades de correlación entre variables, encontramos que no existe tal correlación entre los atributos físicos directamente enlazados con el potencial que marca el registro del jugador, por lo que podemos decir que no existe manera de definir una relación de comportamiento en el que crezca una variable o decrece de manera síncrona.



plt.figure(figsize=(8, 6))

sns.heatmap(prospects_df[['movement_sprint_speed', 'potential']].corr(), annot=True, cmap='coolwarm', linewidths=0.5)

plt.title('Mapa de calor de potencial relativo a la velocidad de sprint')



plt.figure(figsize=(8, 6))

sns.heatmap(prospects_df[['movement_acceleration', 'potential']].corr(), annot=True, cmap='coolwarm', linewidths=0.5)

plt.title('Mapa de calor de potencial relativo a la aceleración de jugador')

Conclusiones

Es de suma importancia decir que el análisis llevado a cabo acerca de la información de los jugadores en la franquicia FIFA nos ha llevado a la conclusión de que importa más la correlación entre las variables categóricas, tales como la liga o la nacionalidad, así como el nivel de ligas y equipos en los que los jugadores juegan cuando de definir el potencial de las jóvenes promesas se trata.

A nivel del videojuego, no se encontró que existieran relaciones directas entre los atributos que se definen para el jugador con el potencial que este tiene, una variable que sirve para proyectar no tiene nada que ver con las habilidades que sus estadísticas marcan y las que el usuario ve.

He de mencionar que las relaciones más importantes que encontré son aquellas en las que las ligas a las que más pertenecen los jugadores jóvenes coinciden con las ligas de mayor rendimiento competitivo en todo el mundo. Dicha sentencia puede favorecer al potencial futuro que tengan estos futbolistas para siguientes entregas de la saga.