

A photograph of a person with long brown hair, seen from the back, looking at a laptop. The laptop screen shows the UAX website with the text 'Desvela la Universidad Online de UAX'. The background is a light-colored wall with a subtle grid pattern.

T04_Análisis Exploratorio de Datos

Román Clemente Jurado

Fundamentos de la Ciencia de Datos

ÍNDICE

Contenido

• ¿QUÉ CARACTERÍSTICAS PREDICEN MEJOR EL OVR DE UN JUGADOR?	3
• ¿CÓMO AFECTA LA POSICIÓN DEL JUGADOR A SU VALORACIÓN PROMEDIO?.....	3
• ¿CÓMO INFLUYEN ATRIBUTOS FÍSICOS Y TÉCNICOS EN EL RENDIMIENTO GENERAL DEL JUGADOR?	3
1. Como dijimos lo primero ya lo hicimos que fue calcular el peso de las correlaciones asique ahora calcularemos el peso VIF:	7
('sho', 24.0546), ('pac', 26.6265), ('shot power', 27.4465), ('jumping', 30.872), ('phy', 34.1425), ('reactions', 42.0407), ('pas', 48.0065), ('dri', 53.5806)	7
2. Ahora normalizamos el VIF y las correlaciones, calculamos las puntuaciones compuestas en base a la formula que previamente explicamos y asociamos las puntuaciones a las variables, y cogemos las 5 características que mejor score tengan:	8
• reactions: 0.7.....	8
• pas: 0.4065.....	8
• composure: 0.4008	8
• phy: 0.382.....	8
• dri: 0.3674.....	8
3.....	8

1. Seleccionar la(s) pregunta(s) que se pueden responder con un modelo de regresión

- **¿Qué características predicen mejor el OVR de un jugador?**

El OVR puede ser modelado como variable dependiente y características como PAC, PHY, y age serían las independientes viendo y prediciendo la linealidad de los datos y su dependencia entre ellos

- **¿Cómo afecta la posición del jugador a su valoración promedio?**
- **¿Cómo influyen atributos físicos y técnicos en el rendimiento general del jugador?**

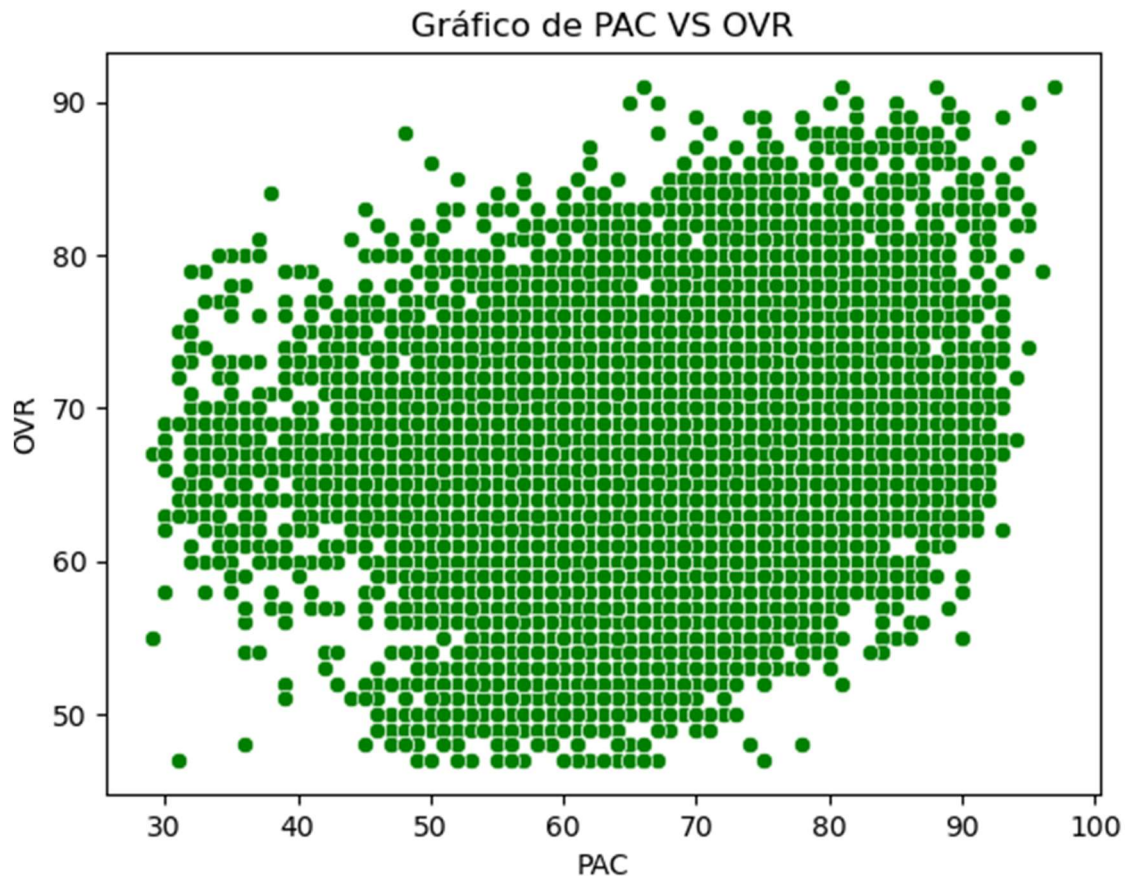
2. Describir el problema, indicando la relación conceptual entre las variables independientes seleccionadas y la variable dependiente

Para predecir OVR

- Variable dependiente o la eje Y situaremos el OVR y en contraposición en el eje X o independiente colocaremos valores como el PAC, PHY, age y position como variable categórica.
- Relación conceptual:
- La velocidad y la fuerza física impactan en el rendimiento general del jugador.
- La edad puede influir negativamente en atributos físicos como velocidad.
- La posición del jugador determina el énfasis en ciertos atributos y que atributos o cosas se potencian, afectando el OVR.

3. Analizar la relación bivariado (por pares) entre la variable dependiente y cada una de las variables independientes

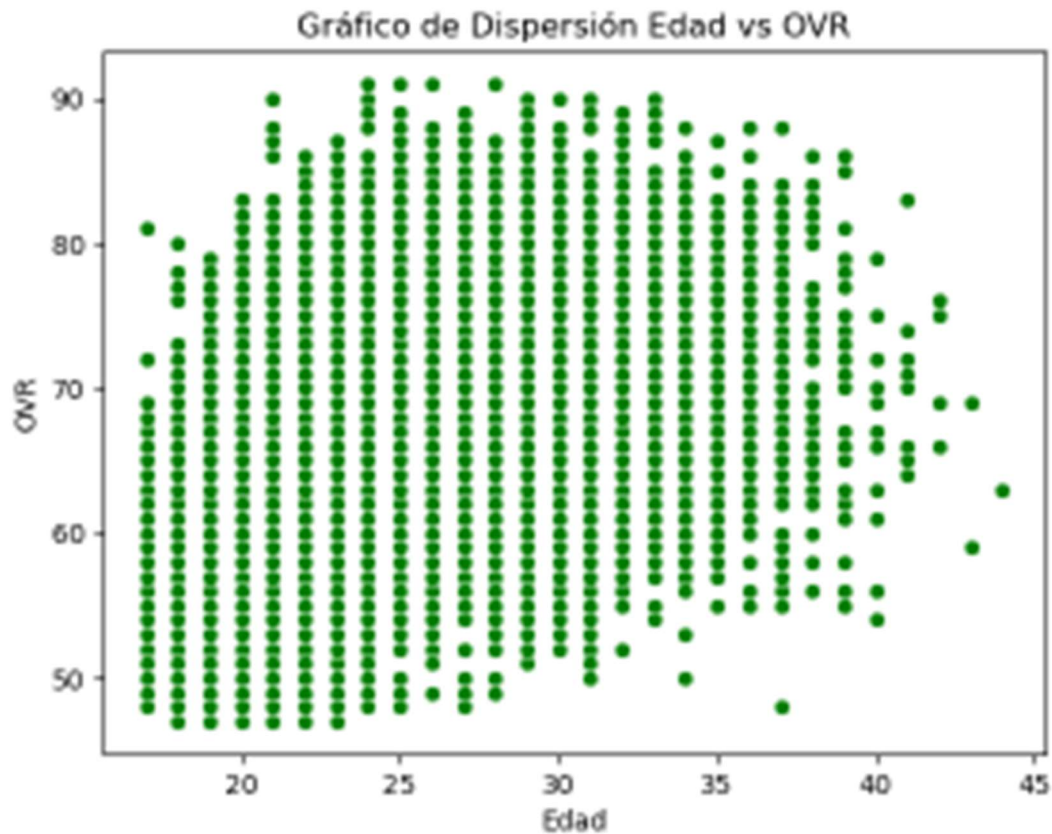
OVR vs PAC



Si observamos la nube de puntos vemos una relación positiva débil entre las variables, por tanto, a medida que PAC aumenta, el OVR tiende a aumentar ligeramente y con un valor de correlación de $r=0.286$

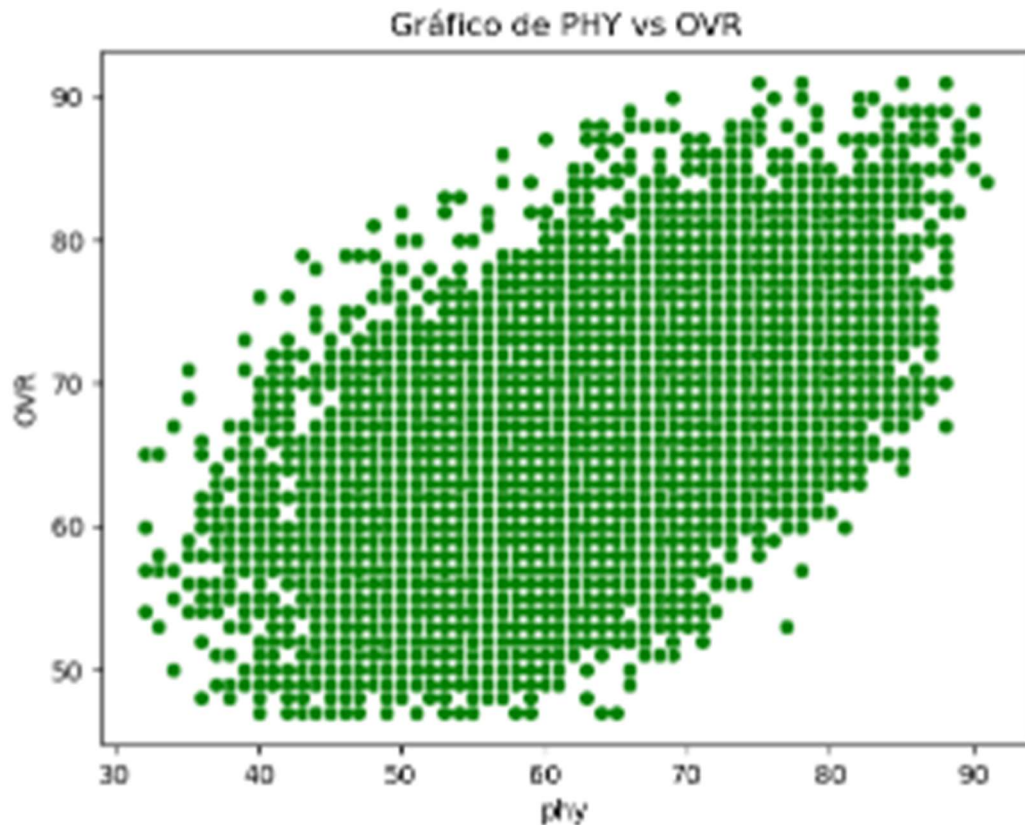
PAC influye en OVR, pero no de manera significativa, esto quiere decir, que los jugadores rápidos no necesariamente tienen una calificación OVR alta, aunque existe cierta tendencia positiva, por tanto, si es influyente dicha capacidad

OVR vs AGE



La dispersión indica que OVR es más elevado en jugadores jóvenes, entorno a los jóvenes de 20-30, y disminuye conforme aumenta la edad aunque la relación parece no lineal, pues la caída no es tan abrupta después de los 30 años al tener un coeficiente de $r=0.372$ indica que la correlación es positiva moderada, por tanto, la edad tiene un impacto moderado en el OVR siendo los jugadores jóvenes los que suelen tener mejores calificaciones, pero la dispersión sugiere que otras variables pueden influir más

OVR vs PHY



La relación entre PHY y OVR es claramente positiva y más consistente que en los otros gráficos, con una correlación de $r=0.567$ → La correlación es positiva moderada a fuerte. Se hace evidente que la fuerza física es una variable importante para predecir el OVR. Los jugadores con mayor PHY tienden a tener mejores calificaciones generales.

Coeficientes de correlación Pearson para medir la fuerza y dirección de la relación.

- Aunque para realizar una buena predicción lo que nos interesa las que tienen las relaciones más fuertes por ello calculamos las correlaciones y filtramos solo obteniendo las más fuertes (superior a 0.5) quedándonos

pas	0.7275608645528306
dri	0.7065191888408587
phy	0.5672655507061907
shot power	0.5560466002182264
vision	0.527320448742756
short passing	0.5383380942332519
long passing	0.525161109997571
reactions	0.8846267563355347
composure	0.6670551711376007

jumping	0.5570435710532776
---------	--------------------

- Aunque he detectado un problema con los datos que no he sabido bien como tratar, al analizar los datos me he dado cuenta que existen una fuerte multicolinealidad entre los mismos, si sometemos al análisis de estos datos mediante el VIF o el factor de inflación de la varianza nos permite cuantificar la intensidad de la multicolinealidad en un análisis de regresión normal de mínimos cuadrados como es nuestro caso y al utilizar la librería statsmodels.stats.outliers_influence podemos analizar dichos datos y por tanto nos ofrece estos valores.

Variable	VIF
pac	64.484313
sho	64.676878
pas	178.348395
dri	369.754455
phy	120.999530
shot power	63.657874
reactions	149.77273
jumping	116.82018

Pero ¿Qué significa esto?

Pues en principio un $VIF > 10$ sugiere una alta multicolinealidad y significa que una variable está correlacionada significativamente con otras variables en el conjunto de datos, por tanto su presencia es redundante si ya hay otra presente, en este conjunto de datos, todas las variables excepto exceden por mucho ese valor lo que denotaría lo que ya mencionamos, lo que normalmente aplicaría en este caso sería buscar la combinación de la mejor categoría de datos que mayor tenga en correlación con el objetivo a buscar en este caso OVR e intentar mantener un perfil VIF bajo en el proceso pero ¿cómo haremos eso?

A la solución que he llegado ha sido crear un algoritmo que haga me permita calcular primero como dijimos ya tenemos las correlaciones y por otro lado calculo las medias de VIF para obtener los que menor valor VIF tienen, entonces eso sumado al peso que asignamos debido a su correlación calculamos esta fórmula:

$$\text{Puntuación} = w1 \cdot \text{Correlación Normalizada} + w2 \cdot (1 - \text{VIF Normalizado})$$

La puntuación se calcula como una combinación lineal ponderada de ambas métricas, donde $w1$ y $w2$ son los pesos que asignaremos a cada métrica según su importancia

Para ello, normalizaremos el VIF promedio cruzado, es decir, convertimos los valores del VIF en una escala entre 0 y 1. Un menor VIF será mejor, por lo que invertiremos los valores y luego normalizaremos la correlación con OVR convirtiendo las correlaciones a valores entre 0 y 1 donde una mayor valor de correlación será mejor.

- Como dijimos lo primero ya lo hicimos que fue calcular el peso de las correlaciones asique ahora calcularemos el peso VIF:

('sho', 24.0546), ('pac', 26.6265), ('shot power', 27.4465), ('jumping', 30.872), ('phy', 34.1425), ('reactions', 42.0407), ('pas', 48.0065), ('dri', 53.5806)

2. Ahora normalizamos el VIF y las correlaciones, calculamos las puntuaciones compuestas en base a la formula que previamente explicamos y asociamos las puntuaciones a las variables, y cogemos las 5 características que mejor score tengan:

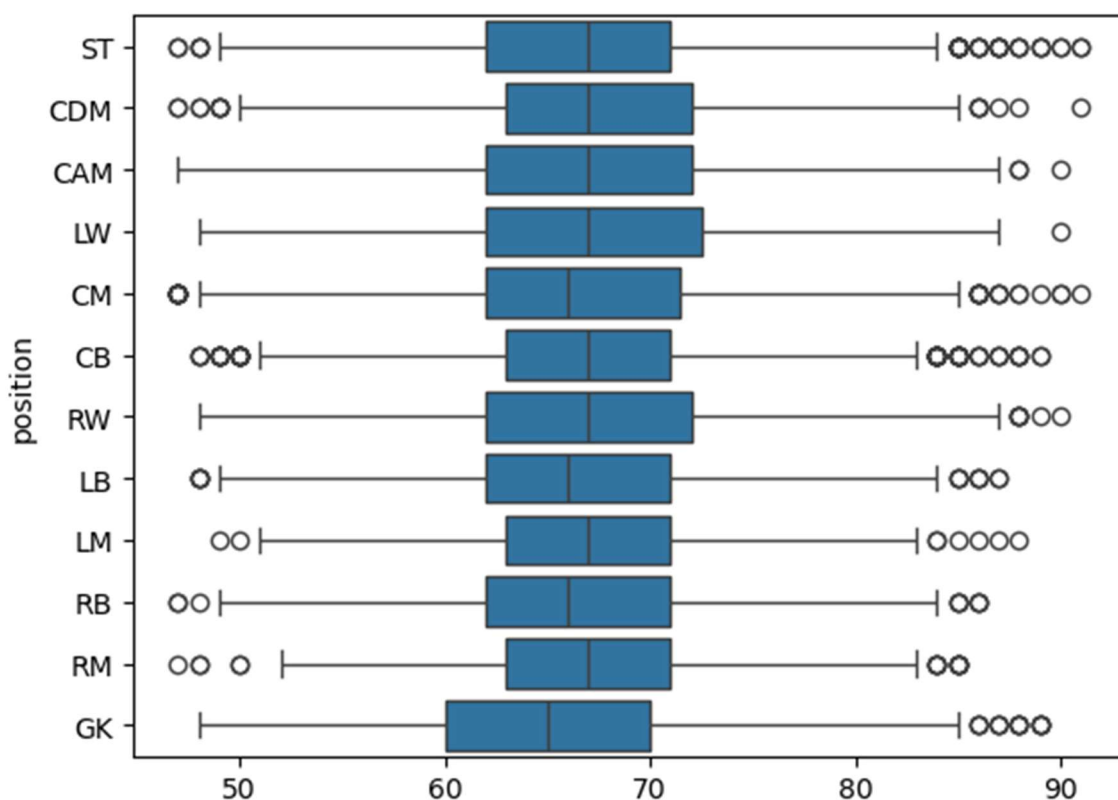
- reactions: 0.7
- pas: 0.4065
- composure: 0.4008
- phy: 0.382
- dri: 0.3674

3.

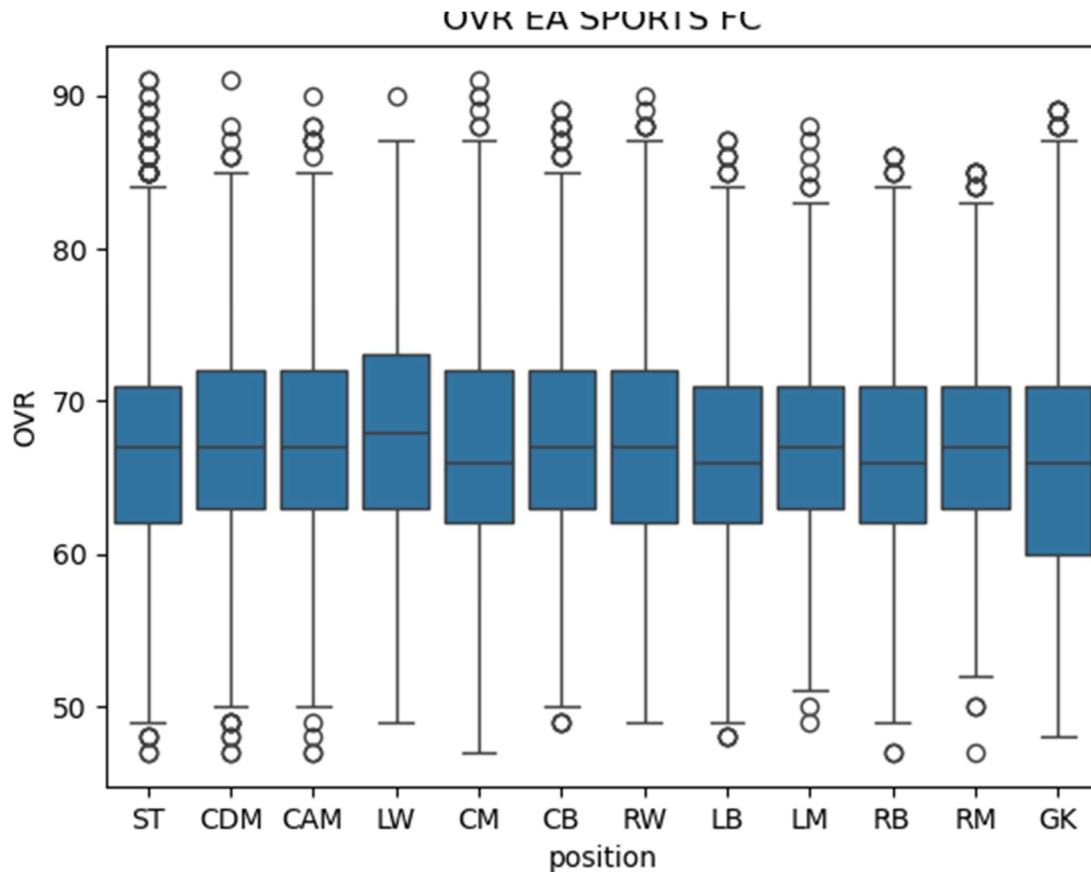
Diagramas de cajas si usamos variables categóricas como position.

Antes de reciclar el diagrama de cajas para ver la posición y su OVR lo que vamos a hacer es limpiar los outliers mediante el rango intercuartílico y estableciendo un límite para que no estropee la regresión lineal

Primero visualizamos los datos que vamos a limpiar:



Calculamos la media y desviación estándar primero y definimos los límites para que no supere por mucho lo establecido nuestro limite que en este caso será de 3, que es un rango asumible y filtraremos los valores dentro de los límites



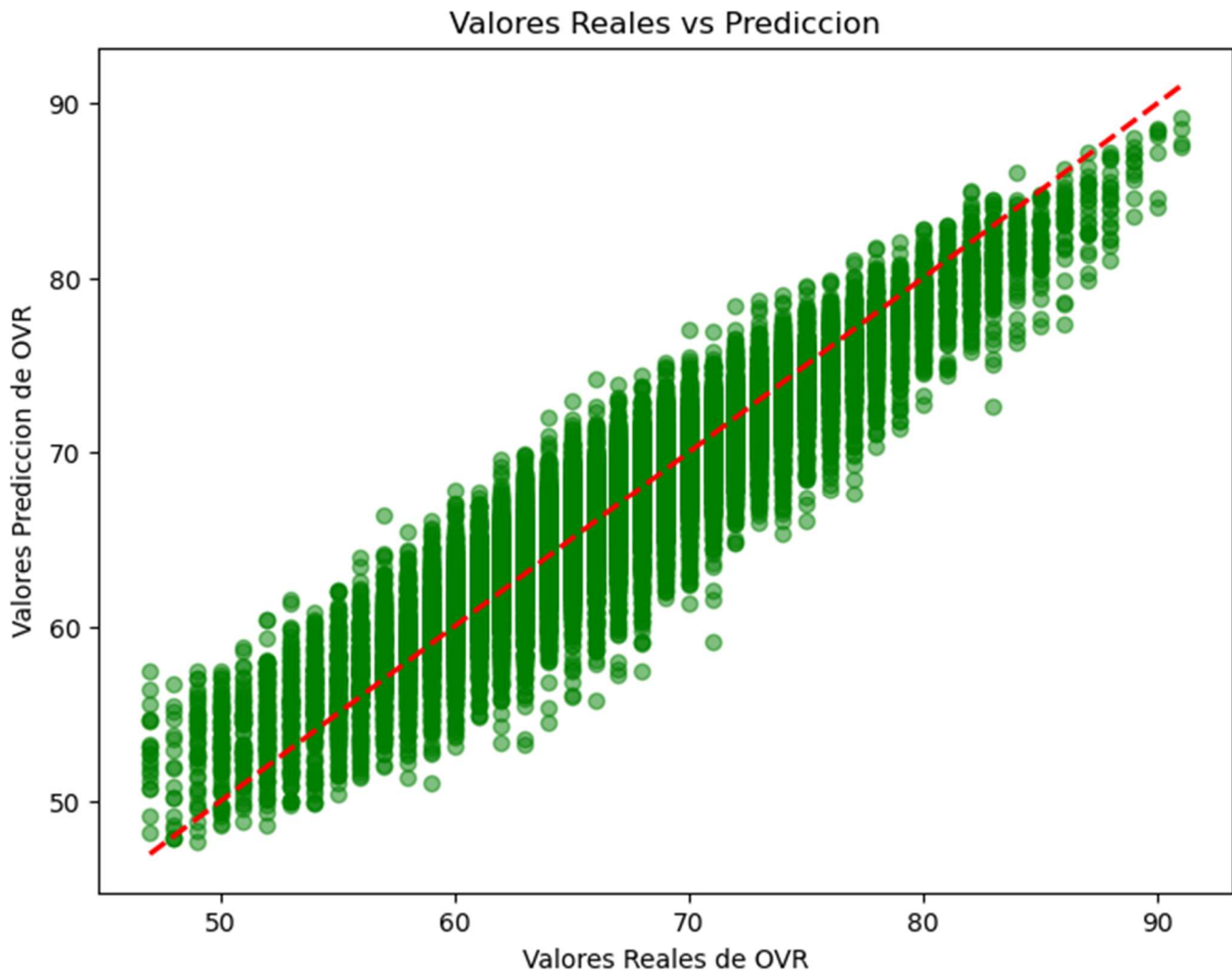
4. Formular el modelo que se ajustará, indicando la transformación que se sugiere para cada variable del modelo

Para empezar lo que vamos a hacer es ajustar y transformar los parámetros que sean necesarios si es que lo son para que sean posibles con la regresión lineal para ello:

En este caso vamos a seguir la formula:

$$\text{OVR} = \beta_0 + \beta_1 \cdot \text{Reactions} + \beta_2 \cdot \text{PAS} + \beta_3 \cdot \text{Composure} + \beta_4 \cdot \text{PHY} + \beta_5 \cdot \text{DRI} + \epsilon$$

5. Ajustar el (los) modelo(s)



- Los puntos muestran la comparación entre los valores reales de OVR y los valores predichos
- La línea roja representa la línea perfecta donde los valores reales y predichos coinciden exactamente
- Cuanto más cerca estén los puntos de la línea roja, mejor es el ajuste del modelo

6. Interpretar resultados

Datos:

- Un MSE o error cuadrático medio de 6.47 indica que, en promedio, las predicciones del modelo tienen un error de alrededor de 6.47 unidades respecto a los valores reales de OVR. Es un error relativamente bajo que como ya he explicado se ha tratado de cierta forma conforme a la naturaleza de los datos
- Un Coeficiente de Determinación o R^2 de 0.864, nos indica que en el 86.4% de la variación en OVR es explicada por las variables independientes seleccionadas, por tanto, nos dice que el modelo no tiene un mal poder predictivo

Coeficientes del modelo:

$$\text{OVR} = 11.88 + 0.38 \cdot \text{reactions} + 0.087 \cdot \text{pas} + 0.048 \cdot \text{composure} + 0.173 \cdot \text{phy} + 0.181 \cdot \text{dri}$$

- El valor 11.88 es el Intercepto y es el valor base de OVR cuando todas las variables iguales a cero sirven como referencia para el ajuste
- 0.38 pertenece a reactions y es la variable con el mayor peso en el modelo. Un aumento de 1 unidad en reactions incrementa OVR en 0.38 unidades, manteniendo el resto constante
- 0.173 corresponde a la fuerza física y también tiene un impacto significativo. Incrementar phy en 1 unidad aumenta OVR en 0.173 unidades
- Por otro lado, el dri = 0.181 es la habilidad de driblar también contribuye de forma importante, con un impacto de 0.181 unidades por cada unidad adicional
- PAS con su valor de 0.087 tiene un efecto más moderado, pero aún positivo
- Y finalmente el 0.048 de composure tiene el coeficiente más pequeño, sigue siendo significativo, indicando que la compostura afecta ligeramente al rendimiento general

7. Conclusiones del análisis (usando lenguaje de negocio)

Conclusiones del análisis:

Variable más influyente:


- La reacción es el predictor más importante para la calificación general. Los jugadores con mejores reflejos y reacciones tienden a tener puntuaciones de OVR significativamente más altas.

Impacto de atributos físicos:

- La fuerza física y la habilidad de driblar son factores clave para predecir el OVR. Esto indica que los jugadores que son fuertes físicamente y buenos en el manejo del balón tienen un mejor rendimiento general.

Atributos técnicos adicionales:

- Aunque con menor peso, el pase y la compostura contribuyen positivamente al OVR. Estos atributos son importantes, pero su impacto es más limitado en comparación con reactions, phy y dri.



UAX

Universidad
Alfonso X el Sabio

GRACIAS

UAX.COM