PREDICCIÓN DEL VALOR DE UN JUGADOR DE FÚTBOL

- Juan Manual Tornero
- Federico Carboni
- Juan Giustina

EMPRESA Y OBJETIVOS

- Empresa: Club de fútbol Argentino.
- Problema: Reemplazar un jugador de fútbol por otro de similares características sin exceder el presupuesto del Club.
- Objetivo: Predecir el valor de un jugador (Y) en base a determinadas características (Xi).

DATASET - FUENTE PRIMARIA

- KAGGLE

FIFA22 OFFICIAL DATASET - ENERO 2022

URL: https://www.kaggle.com/datasets/bryanb/fifa-player-stats-database?select=FIFA22_official_data.csv

DATA WRANGLING / EDA

- Eliminamos columnas que contienen datos irrelevantes (URLs,)
- Transformación de la variable "Value", el cual es nuestro target.
- Nuestra variable target se encuentra en Euros y en algunos casos en miles y otros en millones
- Limpieza de la variable Y --> Eliminamos el signo Euros, unificamos el valor en millones de Euros y cambiamos el tipo de dato: objeto a numérico.
- Eliminamos valores nulos / faltantes.
- Se decide eliminar los atributos correspondientes a los arqueros, ya que son propio de la posición, y desviaría el resultado de los modelos de predicción.

DATA WRANGLING / EDA

- Los jugadores tienen asignados su posicion especifica en la que juegan, a fines de simplificar, los encasillaremos en posicio nes generales (Arquero, defensor, volante y delantero) siguiendo la siguiente foto.

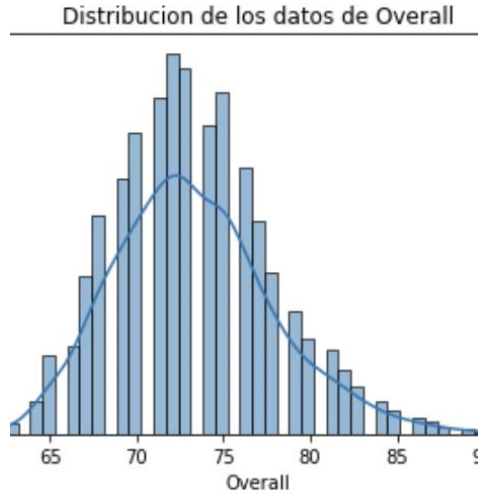
ANÁLISIS DE COMPONENTES PRINCIPALES

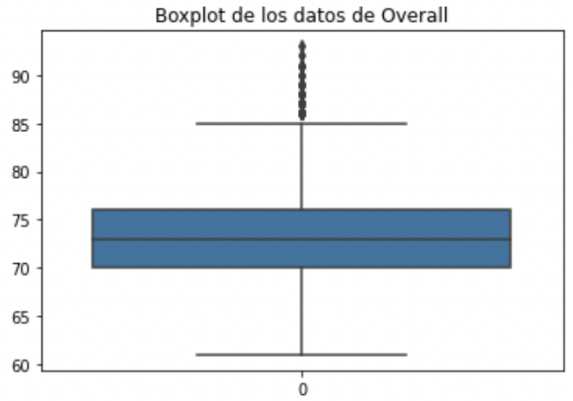
- Variable "Overall": Valoración cuantitativa (numérica) del jugador en términos de habilidades adquiridas.

I. Se realiza .describe() para analizar las principales medidas estadísticas de la variable.

II. Se realiza un histograma para entender si la variable muestra una distribución normal.

III. Se realiza un boxplot para entender si hay outliers presentes que puedan interferir en la conclusión.





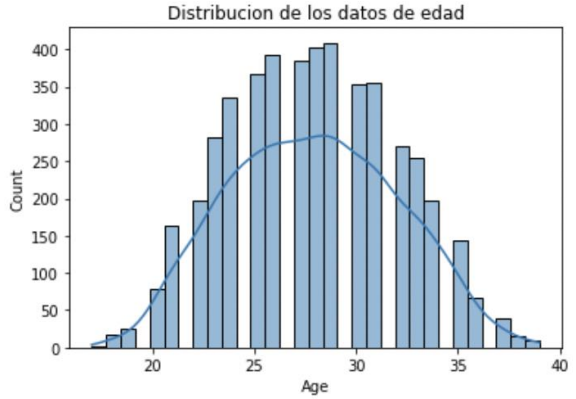
ANÁLISIS DE COMPONENTES PRINCIPALES

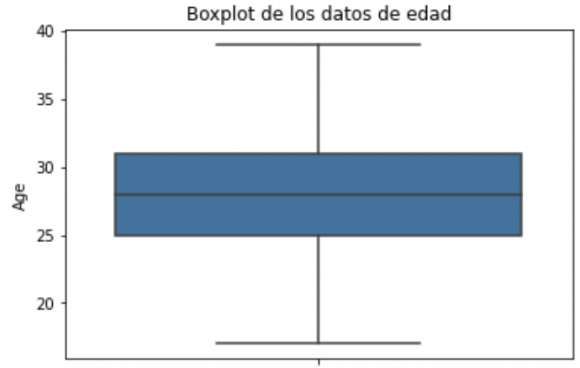
- Variable "Age": Valor cuantitativo (numérico) que representa la edad del jugador.

I. Se realiza .describe() para analizar las principales medidas estadísticas de la variable.

II. Se realiza un histograma para entender si la variable muestra una distribución normal.

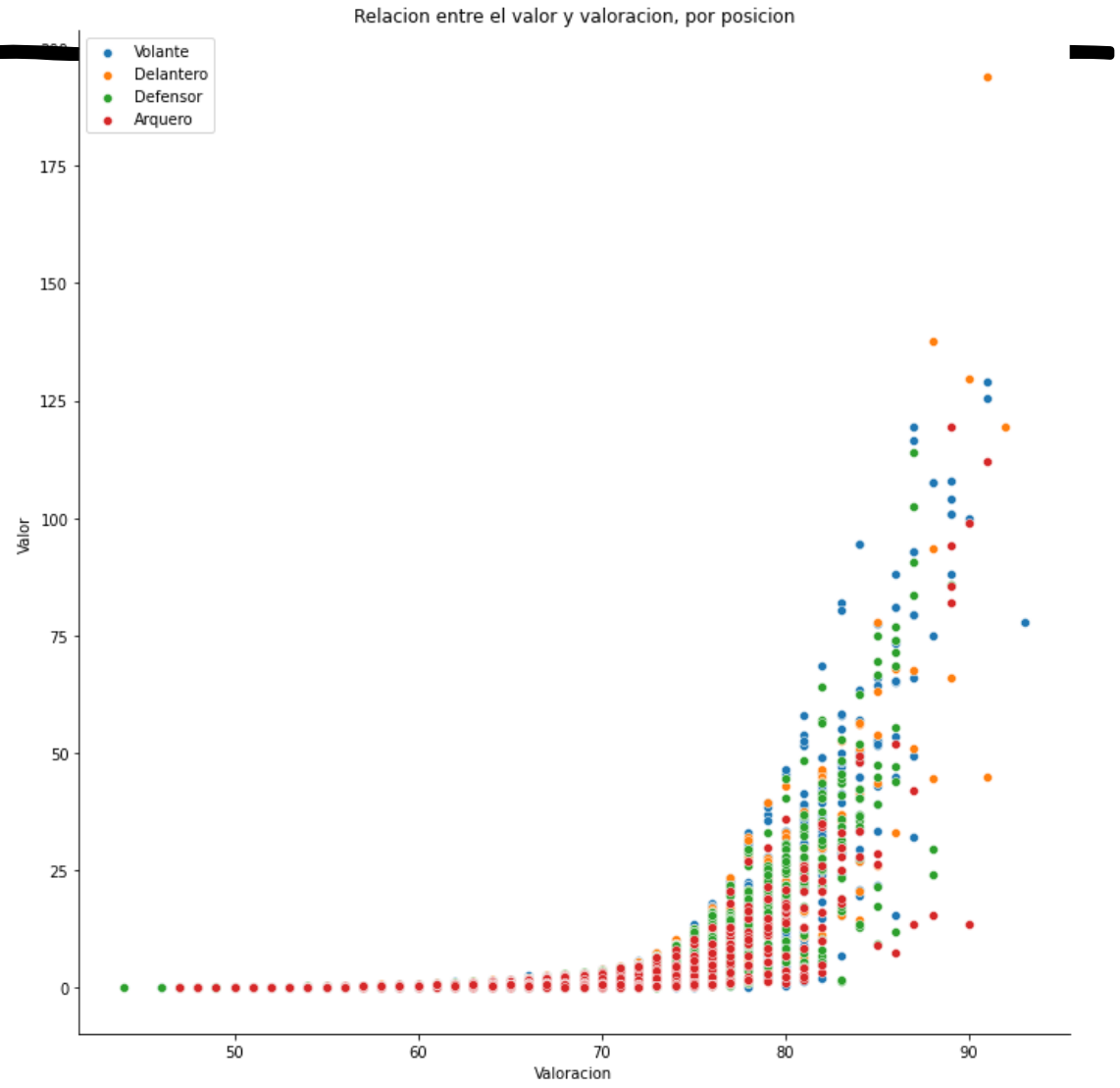
III. Se realiza un boxplot para entender si hay outliers presentes que puedan interferir en la conclusión.





RELACION ENTRE VARIABLES

- Entre todos los atributos con los que contamos, pasaremos a verificar si se observa relacion entre la valoracion total del jugador y su correspondiente valor de mercado, segmentado por posicion.
- Se observa que hay una relacion positiva entre el valor y la valoracion, y que el arquero es la posición menos valorizada del mercado.



VARIABLES ELEGIDAS PARA EL MODELO

- Realizamos otro filtrado de variables, eliminando datos como la nacionalidad, el club, el año que se unió al club, pais donde juega, entre otros. No las consideramos relevantes. A su vez, eliminamos a los arqueros. Creemos que ellos deberían tener un modelo propio.
- Las variables a utilizar para predecir el modelo se nombran en la diapositiva siguiente. Un breve comentario de ellas, es que hemos decidido avanzar con variables cuantitativas objetivas sobre la valoración de los jugadores en determinadas habilidades, que van de 0 a 100, siendo 0 la peor calificación y 100 la mejor. Hemos tenido en cuenta la edad del jugador, y su potencial a futuro. Y para terminar, otro condimento que podría inferir en el resultado es la liga donde el jugador se desempeña. Creemos que esta valoración es subjetiva, sin embargo, la consideramos que esta implícita en la variable Overall de los jugadores.

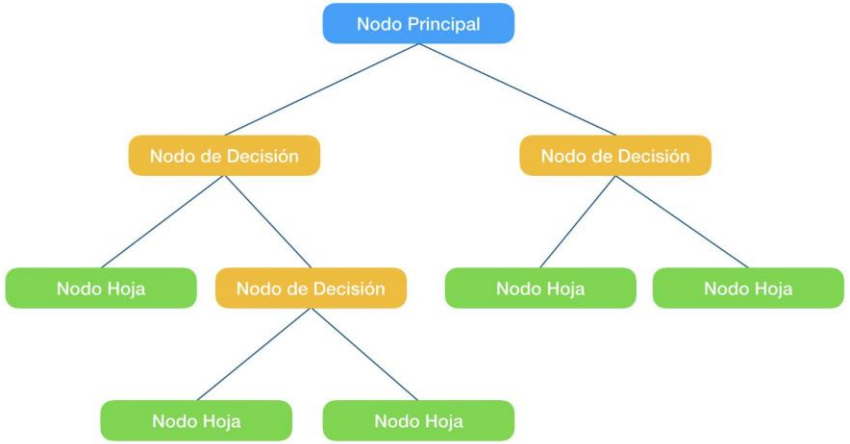
VARIABLES ELEGIDAS PARA EL MODELO

- Edad
- Overall
- Potencial
- Reputación Internacional
- Crossing
- Definicion
- Pase corto
- Volea
- Dribbling
- Curva
- Pase largo
- Control
- Aceleración
- Agilidad
- Reaccion
- Balanceo
- Fuerza de tiro
- Salto
- Stamina
- Fuerza
- Tiro largo
- Agresion
- Intercepcion
- Posicionamiento
- Vision
- Tackle parado
- Tackle con deslizamiento

ALGORITMOS APLICADOS - ARBOL DE DECISIÓN

Indicadores de performance del modelo:

Raíz del error cuadrático medio - Train: 2.615 / Raíz del error cuadrático medio - Test : 4.533
Ajuste del valor de R cuadrado - Train: 0.956 / Ajuste del valor de R cuadrado - Test : 0.916



Conclusiones:

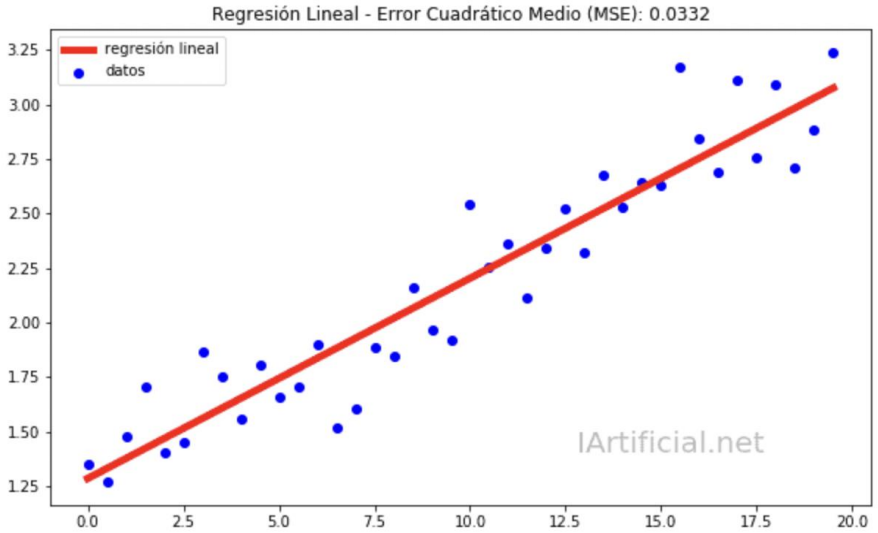
Con una profundidad menor, =3, se obtienen buenos resultados tanto en el conjunto de train y test. Estos resultados son parecidos, lo que indica que el modelo no sufre de varianza.

Con una profundidad de 15, se obtiene un R2 = 1, por lo que se confirma el overfitting.

ALGORITMOS APLICADOS - REGRESIÓN LINEAL

Indicadores de performance del modelo:

Raíz del error cuadrático medio - Train: 7.402 / Raíz del error cuadrático medio - Test : 9.632
Ajuste del valor de R cuadrado - Train: 0.648 / Ajuste del valor de R cuadrado - Test : 0.62



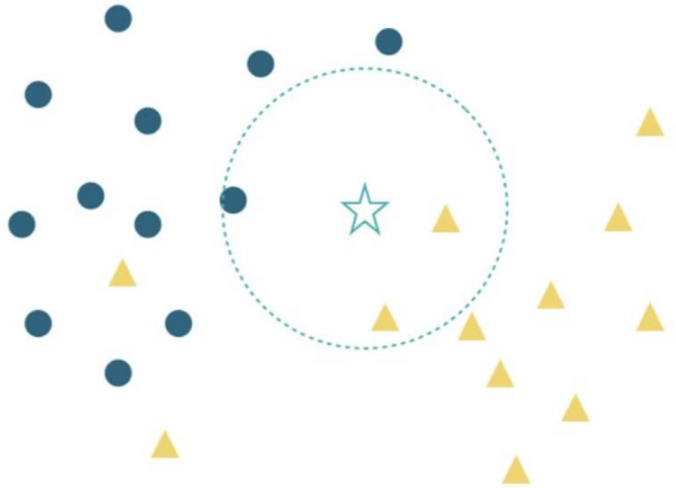
Conclusiones:

Podemos observar que no se obtienen buenos resultados, dado que el modelo representa solo al 62% de los datos en el conjunto de test. Se podría realizar una validación cruzada para mejorar el modelo

ALGORITMOS APLICADOS - KNN (VECINOS CERCANOS)

Indicadores de performance del modelo:

Raíz del error cuadrático medio - Train: 3.029 / Raíz del error cuadrático medio - Test : 5.034
Ajuste del valor de R cuadrado - Train: 0.857 / Ajuste del valor de R cuadrado - Test : 0.678



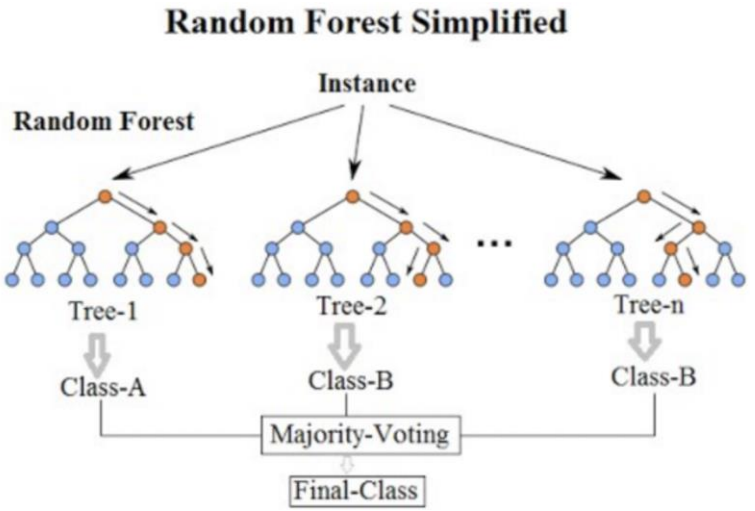
Conclusiones:

A partir de prueba y error, se comprueba que los mejores resultados se obtienen con una cantidad de vecinos entre 5 y 10. Sin embargo, observamos una diferencia consistente entre las métricas del conjunto de datos de entrenamiento y de test, lo que nos dice que existe una varianza alta.

ALGORITMOS APLICADOS - RANDOM FOREST

Indicadores de performance del modelo:

Raíz del error cuadrático medio - Train: 2.264 / Raíz del error cuadrático medio - Test : 3.092
Ajuste del valor de R cuadrado - Train: 0.92 / Ajuste del valor de R cuadrado - Test : 0.878



Conclusiones:

Con el modelo de random forrest se han mejorado los resultados del modelo de árbol de decisión. Mantuvimos la profundidad del árbol de decisión realizado previamente, y variamos la cantidad de áboles a realizar, sin tener demasiadas diferencias en el resultado.

CONCLUSIONES

- A partir de los resultados obtenidos, consideramos al random forrest como modelo predictivo. Por un lado, el modelo represento al 87% de los datos del conjunto de entrenamiento, y, más importante, ha brindado un error cuadrático medio de 2.26, traducido a la unidad de nuestra variable target significa que el modelo otorga un resultado que puede variar entre 2.26 millones de Euros más o menos de su valor real de mercado. En el mercado los valores son negociables, asique consideramos que un jugador podría ser negociado con una varianza de ese valor sobre su valor real.