I. Introducción

* El objeto del estudio
Es predecir la valoración overall de un jugador a partir de unas características físicas y técnicas, esto es interesante por cuanto podría darnos idea de la valoración global del jugador teniendo en cuenta algunas de las características que son medibles.

* Los objetivos y alcance del proyecto
Se extienden a la determinación de la variable overall a partir de algunas de las variables del dataset, concretamente tras un análisis exploratorio de los datos y machine learning las variables que utilizamos son las variables 'potential','value_eur','wage_eur','age', 'pace', 'shooting', 'passing', 'dribbling', 'defending' y 'physic'.

Objetivo del Análisis: El propósito del análisis es demostrar o no la viabilidad de utilizar datos de los juegos FIFA para la predicción de talento futbolístico mediante su coeficiente overall.

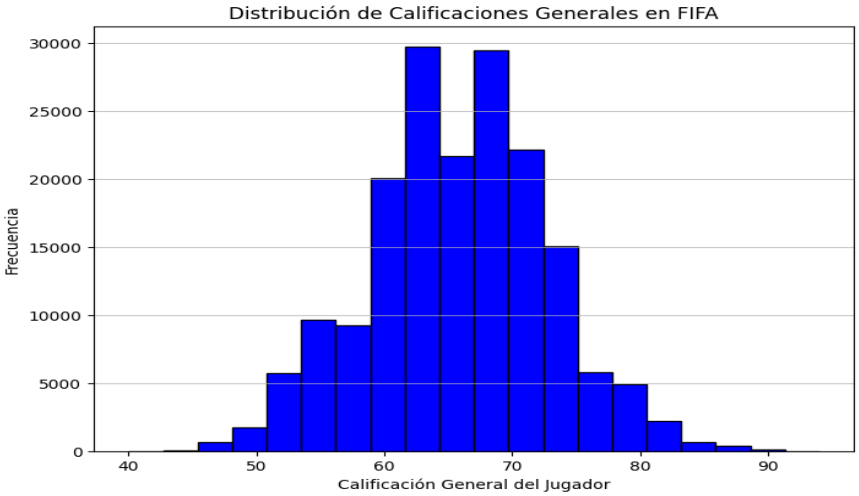
II. Dataset

* Descripción del dataset utilizado (origen, tamaño, variables, etc.)
El dataset utilizado es un dataset de 180021 filas y 109 columnas entre las que contamos con variables integer,float y object.
Los datos de los juegos FIFA de 2015-2024 son obtenidos de la pagina web de datasets kaggle.
Los datos en formato csv son descargados de kaggle en un archivo de aprox 180.021 lineasx109 columnas.

* Análisis exploratorio de los datos (EDA)
El análisis exploratorio de los datos (EDA) se enfoca en la comprensión del dataset y de las variables que en el intervienen.
Se analiza el dataset desde el punto de vista de análisis univariante y multivariante, obteniéndose un entendimiento profundo del mismo del cual se derivan acciones como la determinación de las variables mas importantes.

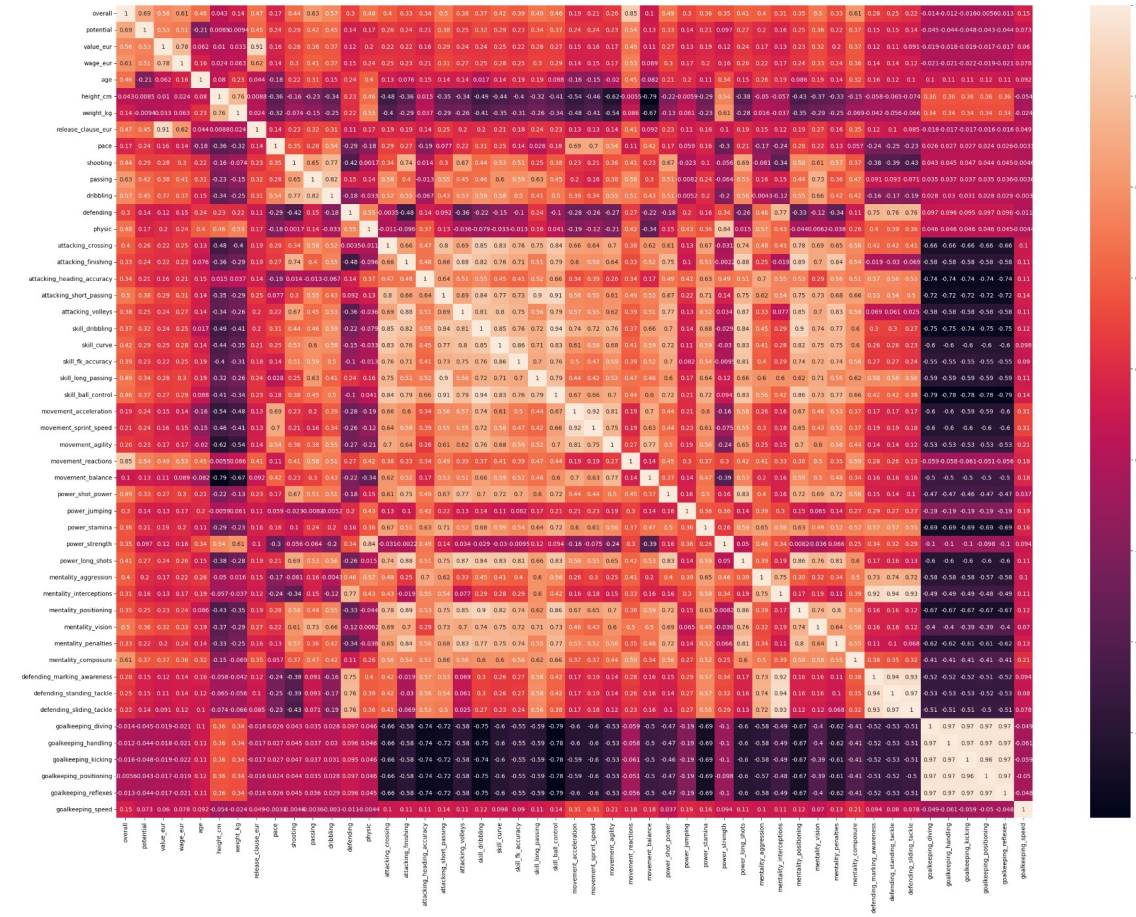
1. Análisis Univariado

- Distribuciones de Variables: Se determinan las variables mas importantes como el valor overall(general) y el potencial de cada jugador, ademas de la edad, valor de mercado y su salario. También se utilizan otras variables secundarias categóricas y numéricas de carácter técnico como el pase(passing),regate(dribbling),defensa(defending),tiro(shooting),físico, (physic) y velocidad-ritmo(pace).
- Medidas Estadísticas: Para las comparativas se usa la mediana ya que es la medida que mejor se ajusta a todas las variables, aunque se han revisado la mediana y la moda para dos variables en concreto los salarios y el valor del jugador.
- Identificación de Outliers: Al realizar el análisis univariado se detectan outliers en los datos sobretodo en la variable valor de mercado y salarios.



2. Análisis Bivariado

- Relaciones entre Variables: Se realiza el análisis de las variables, sin las variables objeto y sin algunas variables que no aportan nada al estudio.
- Correlaciones: Respecto a las correlaciones, se estudian las mismas tanto para las variables primarias como las secundarias con especial énfasis en la relación del overall y el valor de mercado y el potencial con las variables principales y secundarias de carácter técnico.
- Se calculan las correlaciones entre todas las variables, viéndose que existe una fuerte colinealidad entre algunas de ellas.



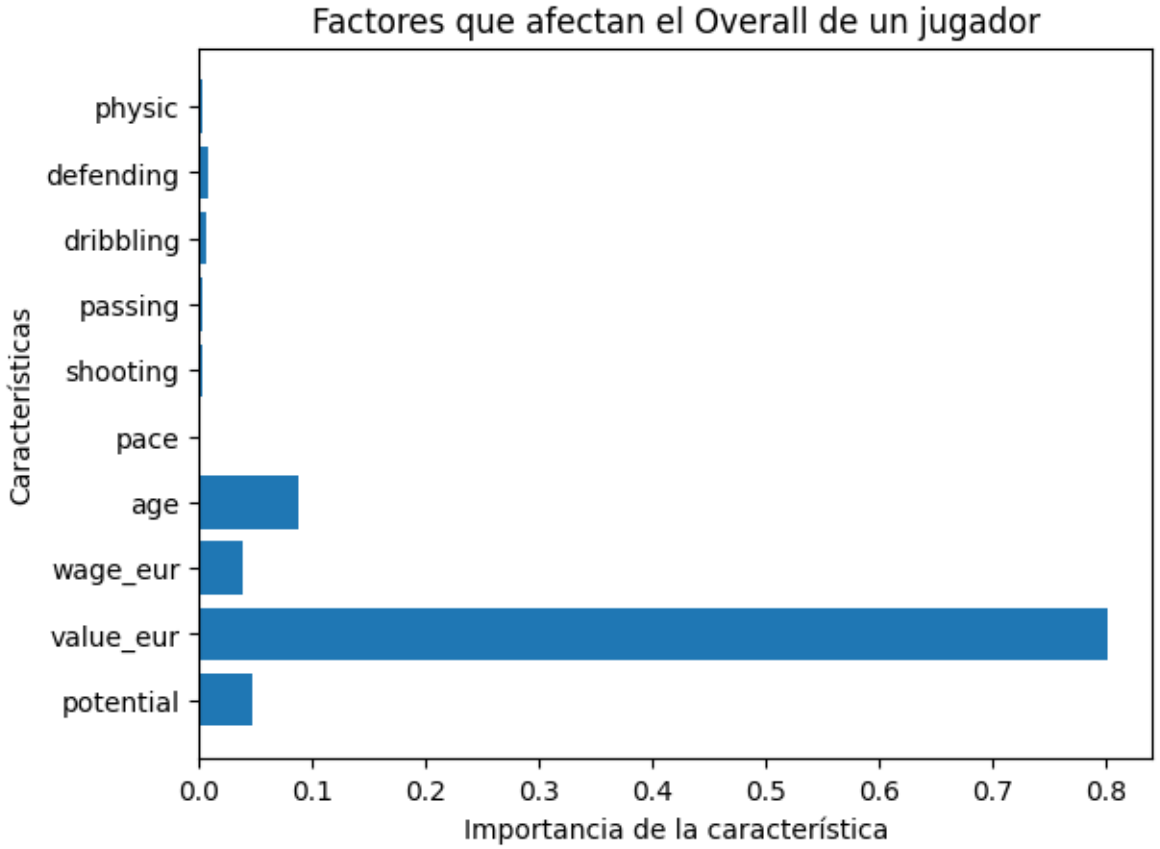
III. Preprocesamiento de los datos

* Verificación de la calidad de los datos

Los datos están distribuidos en varios tipos como integer,float y object de forma correcta a su vez contamos con valores faltantes para algunas variables como value_eur y wage_eur.
No se detectan registros duplicados que puedan distorsionar el modelo.

* Decisiones, imputaciones y transformación de variables

Respecto a los datos efectuamos una primera criba de variables por no considerarlas interesantes, a continuación con un feature importance sobre el total del resto de variables, nos quedamos con 'potential','value_eur','wage_eur','age', 'pace', 'shooting', 'passing', 'dribbling', 'defending' y 'physic'.
Tratamos los faltantes aplicando la mediana por considerarla la mas conveniente.



IV. Modelado

* Entrenamiento de modelos supervisados/no supervisados

Nos centramos en los modelos supervisados por tratarse de un dataset con un target conocido, este target es el overall.
Este modelo usa un Random Forest Regressor, que es un método supervisado basado en árboles de decisión, capaz de identificar qué características tienen mayor impacto en la predicción del overall de los jugadores.
Para predecir la valoración global de un jugador a partir de sus características físicas y técnicas, hay varios enfoques de aprendizaje automático que pueden ser muy útiles. Algunos de los más efectivos incluyen:

1. Regresión Lineal o Regresión Ridge: Si la valoración global se expresa como una variable

continua (por ejemplo, una puntuación de 0 a 100), los modelos de regresión pueden ser una opción simple y efectiva para determinar cómo influyen diferentes características en la valoración.

2. Árboles de Decisión y Random Forest: Estos modelos pueden capturar relaciones no lineales entre características y valoración, permitiendo una interpretación más clara de cuáles factores pesan más en la evaluación.
3. Gradient Boosting Machines (XGBoost, LightGBM, CatBoost): Son modelos más avanzados que optimizan el proceso de aprendizaje y suelen dar excelentes resultados en problemas de predicción con datos tabulares.
4. Redes Neuronales Artificiales: Si dispones de una gran cantidad de datos y las relaciones entre variables son complejas, las redes neuronales pueden captar patrones profundos. Un modelo como una red neuronal profunda (Deep Learning) puede ser eficaz si los datos son suficientes para evitar el sobreajuste.
5. K-Nearest Neighbors (KNN): Este método evalúa la valoración de un jugador en función de los jugadores más similares en el conjunto de datos, lo que puede ser útil si los datos de jugadores anteriores son representativos.
6. Support Vector Machines (SVM): Si la valoración global del jugador se puede clasificar en categorías (por ejemplo, "bajo", "medio", "alto"), un modelo de clasificación como SVM podría ser apropiado.

*** Evaluación de los diferentes modelos e iteraciones.**
Empezamos probando un Random Forest para las 64 variables para poder detectar las feature importances y desarrollar el resto de acciones a partir de ese punto.
Probamos un total de 8 modelos un modelo de regresión lineal, un ridge regression, un ramdonforest, un Xgboost, un Gradient boosting, un SVR, un LightGBM y un Gridsearch de randomforest midiendo su MAE.

Modelo	Error de predicción	Tiempo de ejecución
Regresión Lineal	1.8178	55 sg
Ridge Regression	1.7403	1 sg
RandomForest	0.4339	6 min 54 sg
XGBoost	0.6148	0.4 sg
Gradient Boosting	0.6358	1 min 6 sg
SVR (kernel='rbf', C=100, gamma=0.1)	128 min	STOP
LightGBM	0.6405	0.6 sg
RandomForest con GridSearch	0.4434	5 min 32 sg

RandomForest con GridSearch (0.4434) y RandomForest sin ajuste (0.4339) tienen los errores más

bajos, seguidos por XGBoost (0.6148), Gradient Boosting (0.6358) y LightGBM (0.6405). SVR ha sido extremadamente lento (128 min) y fue detenido, lo que puede indicar que no es una opción eficiente en este caso.

La regresión lineal y Ridge tienen errores más altos, lo que puede significar que el problema no es lineal y necesita modelos más complejos.

Si buscamos equilibrio entre precisión y tiempo de ejecución, XGBoost y LightGBM parecen buenas opciones, ya que tienen errores relativamente bajos y tiempos rápidos. RandomForest con GridSearch también tiene buen desempeño, aunque toma más tiempo (5 min 32 sg).

A través de una feature importance se determinan las variables a tener en cuenta que son con las que realizamos las pruebas restantes, estas variables son 'potential','value_eur','age', 'release_clause_eur', 'movement_reactions', 'wage_eur', 'defending', 'skill_ball_control', 'mentality_positioning', 'mentality_interceptions'.
Los resultados que se obtienen son los siguientes:

Modelo	Error de predicción	Tiempo de ejecución
RandomForest	0.4370	1 min 18 sg
XGBoost	0.6371	0.2 sg
Gradient Boosting	0.6524	1 min 1 sg
LightGBM	0.6527	0.7 sg
RandomForest con GridSearch	0.6764	1 min 8 sg

Parece que el modelo RandomForest sigue siendo el que tiene el error más bajo (0.4370), y además ha reducido significativamente su tiempo de ejecución en comparación con la iteración anterior. Por otro lado, XGBoost sigue teniendo un tiempo de ejecución extremadamente rápido (0.2 sg), aunque con un error algo mayor (0.6371).
Viendo que el error no mejora sustancialmente nos inclinamos por coger variables que siendo significativas son mas fácilmente interpretables tales como 'potential','value_eur','wage_eur','age', 'pace', 'shooting', 'passing', 'dribbling', 'defending' y 'physic'.

Modelo	Error de predicción	Tiempo de ejecución
RandomForest	0.4376	1 min 6 sg
XGBoost	0.6148	0.4 sg
Gradient Boosting	0.6358	1 min 6 sg
LightGBM	0.6405	0.5 sg
RandomForest con GridSearch	0.6765	1 min 8 sg

Parece que el RandomForest sigue teniendo el menor error de predicción (0.4376) con un tiempo de ejecución razonable. XGBoost sigue siendo el más rápido (0.4 sg), aunque con un error más alto. Mientras tanto, Gradient Boosting y LightGBM tienen errores similares, pero LightGBM es ligeramente más rápido.
De ahí que elegimos como variables definitivas las variables 'potential','value_eur','wage_eur','age', 'pace', 'shooting', 'passing', 'dribbling', 'defending', 'physic'.

*** Selección e interpretación del modelo final.**
El modelo seleccionado es un un Random Forest de MAE (0.4376) 1min 6sg con un coste computacional moderado comparado con otros modelos como el modelo SVR que se tuvo que detener sin finalizarlo tras 128min de proceso.

La selección del modelo Random Forest se basa en su capacidad para:

- Manejo de datos no lineales: Puede capturar relaciones complejas que la regresión lineal no detecta.
- Reducción del sobre ajuste: Promedia el resultado de múltiples árboles, lo que reduce el riesgo de aprender patrones espurios.
- Importancia de las variables: Permite evaluar qué características tienen más peso en la predicción.
- Escalabilidad: Funciona bien con grandes volúmenes de datos sin perder eficiencia.

El modelo presenta un MAE de 0.4376 pero presenta un MAE de entrenamiento de 0.1677 lo cual nos lleva a interpretar que existe overfitting, el MAE en entrenamiento es bajo (0.1677), lo que sugiere que el modelo aprende bien en los datos de entrenamiento.

Sin embargo, el MAE en prueba es significativamente mayor (0.4376), lo que indica que el modelo no generaliza bien en datos nuevos.
Para mejorarlo se trata de reducir la complejidad del modelo bajando max_depth en el Gradient Boosting para evitar árboles demasiado profundos.

Reduciendo el n_estimators para evitar sobre aprendizaje en el conjunto de entrenamiento lo cual no arroja mejoras significativas.

Usando un learning_rate más bajo para suavizar el aprendizaje, pero ninguna de estas técnicas ofrece resultados óptimos.

También usamos el cross_val_score() para verificar si el modelo es estable en diferentes particiones de datos y con un valor de 0.4508 parece que es así.

Por tanto finalmente nos quedamos con ese MAE del Randomforest de 0.4376 que es el que nos permite centrarnos en las variables que queremos manejar a un MAE aceptable.

V. Predicción y resultados finales

*** Descripción de la solución final y su impacto en el negocio**
La solución final es un RandomForest para generar predicciones precisas basadas en datos históricos.
Sus características principales incluyen:

- Precisión y estabilidad: La combinación de múltiples árboles reduce el sobre ajuste y mejora la generalización.

- Análisis de importancia de variables: Permite identificar qué factores tienen mayor peso en la toma de decisiones.
- Automatización y escalabilidad: Puede adaptarse a grandes volúmenes de datos y actualizarse con nueva información.
- Optimización mediante hiperparámetros: Ajustes en n_estimators, max_depth y max_features normalmente mejoran la performance del modelo aunque no es el caso de este dataset.

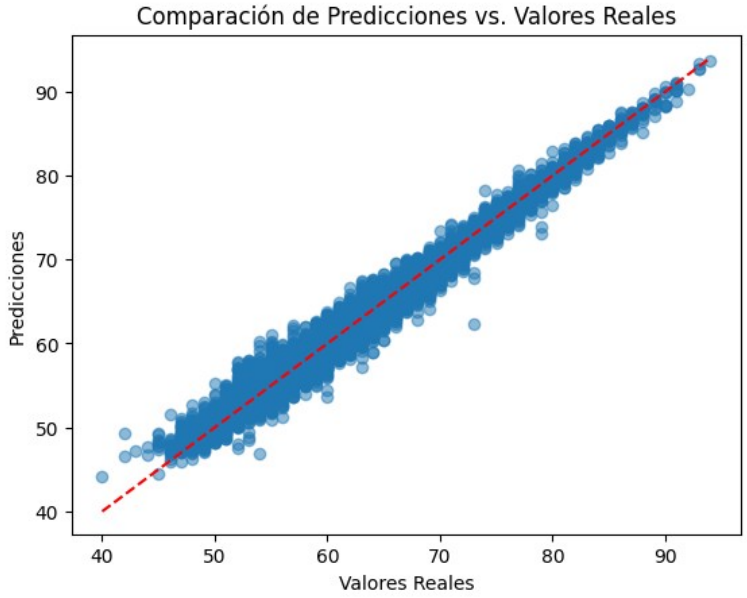
Impacto en el negocio

Implementar esta solución ofrece múltiples beneficios estratégicos:

1. Mejor toma de decisiones
    - El club podrá prever tendencias de los jugadores con mayor precisión, ajustando estrategias comerciales y operativas.
2. Reducción de costos y optimización de recursos
    - Evita gastos innecesarios al mejorar la predicción del overall del jugador y siendo un elemento efectivo en la determinación de la idoneidad de un jugador así como reduciendo errores de estimación.
3. Ventaja competitiva
    - Permite adelantarse a cambios en el mercado y reaccionar con agilidad, contando con una herramienta que permite diferenciarse de la competencia.

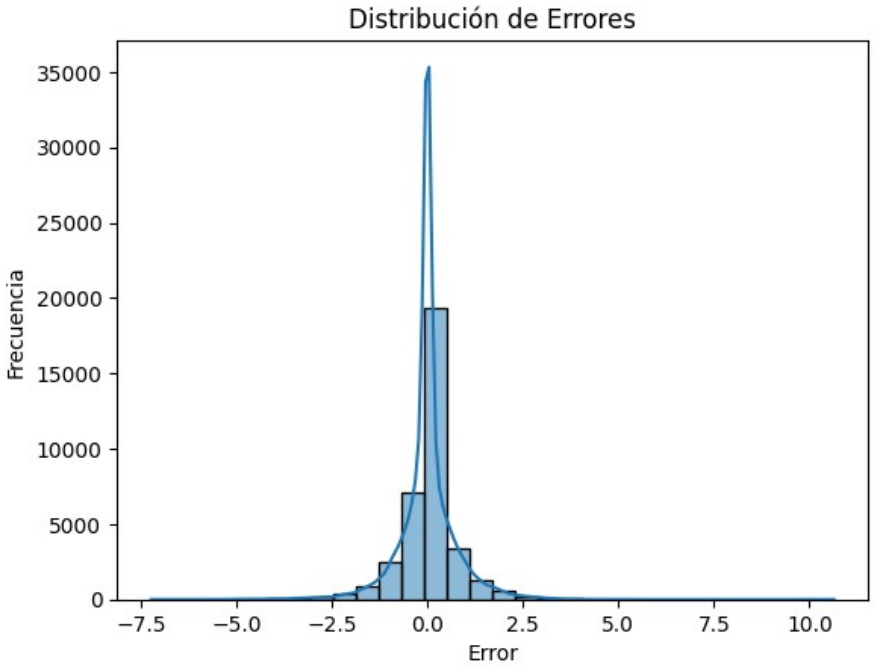
*** Visualización de los resultados finales y predicciones**

Este gráfico de dispersión nos permite ver qué tan cerca están las predicciones de los valores reales.



Histograma de Errores

Este gráfico muestra la distribución de los errores (diferencia entre valores reales y predichos).



VI. Conclusiones y pasos futuros

* Análisis de los resultados y fortalezas/debilidades del proyecto

La implementación de Random Forest para regresión ha permitido obtener una solución robusta, pero es esencial evaluar qué tan bien ha funcionado, identificando fortalezas y áreas de mejora.

* Propuesta de futuras mejoras y optimizaciones

Fortalezas del Proyecto

1. Alta precisión y estabilidad
    - Random Forest reduce el sobreajuste al combinar múltiples árboles.
    - Funciona bien con datos con estructuras no lineales.
2. Capacidad de manejar datos con muchas variables
    - Al analizar la importancia de características, podemos determinar qué factores tienen mayor influencia.
3. Escalabilidad
    - Se puede aplicar a grandes volúmenes de datos sin comprometer el rendimiento.
4. Flexibilidad en la interpretación
    - Se pueden ajustar hiperparámetros como max_depth, n_estimators y max_features para optimizar el rendimiento.

Debilidades y Áreas de Mejora

1. Alta demanda computacional
    - Puede ser más lento en comparación con modelos lineales simples.
    - Si el dataset es muy grande, se pueden considerar métodos de optimización como

reducir el número de árboles o usar técnicas de muestreo.

2. Menor interpretabilidad
    - Aunque permite evaluar la importancia de las variables, no proporciona coeficientes directos como una regresión lineal.
3. Riesgo de sobreajuste en ciertos escenarios
    - Si se usa demasiada profundidad en los árboles o una gran cantidad de características, el modelo podría memorizar los datos en lugar de generalizar.
4. Posible sensibilidad a datos ruidosos
    - Si el dataset tiene muchas variables irrelevantes o datos inconsistentes, la calidad de las predicciones puede verse afectada.
    - Se recomienda hacer una buena selección de características antes de entrenar el modelo.

Conclusión y Próximos Pasos

El modelo ofrece una solución sólida para la predicción de valores con un buen equilibrio entre precisión y robustez. Sin embargo, hay margen de mejora en interpretabilidad y optimización. Dependiendo del caso de uso, se pueden explorar alternativas como:

- Optimización del modelo ajustando hiperparámetros.
- Uso de modelos más interpretables si la explicabilidad es clave.
- Mejor selección de características para reducir ruido en los datos.

VII. Referencias

- Arsene Wenger: desarrollando al jugador del futuro
- Incrementar-la-competitividad-mundial-Un-analisis-del-ecosistema-de-desarrollo-del-talento.pdf
- Detección de la capacidad del talento en el jugador de fútbol
- Transición de talentos | FIFA Publications
- How data analysis helps football clubs make better signings
- Microsoft Word - informeFinal_segundoBorrador.docx
- 05) Detección del talento - FIFA Training Centre
- Find Open Datasets and Machine Learning Projects | Kaggle
- Visual Studio Code - Code Editing. Redefined
- pandas - Python Data Analysis Library
- Matplotlib — Visualization with Python
- seaborn: statistical data visualization — seaborn 0.13.2 documentation
- Thesis Template