Valores de Jugadores de Futbol

¿Se puede Predecir el valor de un Jugador de Futbol?

AUTOR: Alessandrini Pablo.

AGENDA

- 01 Contexto y Audiencia
- 02 Hipótesis/Preguntas de Interés/Metadata
- 03 EDA
- 04 Creación de Modelos de Machine Learning
- 05 Optimización de Modelo y Visualización de Predicciones.

CONTEXTO Y **AUDIENCIA**

Contexto

El fútbol es del deporte más popular del mundo. Teniendo un universo analítico, y con visión de negocio, es importante analizar variables, tales sean “Valor Actual”, “Apariciones” y “Edad”.

1. **Edades de los jugadores.**
2. **Altura de los jugadores**
3. **Cantidad de apariciones.**

El dataset, que se utilizará en el proyecto, proviene de una base de datos de cualidades y estadísticas de jugadores de fútbol de la mayoría de los equipos del mundo, correspondientes a los años 2021 y 2022. Entre los datos disponibles se incluyen las posiciones en el campo de juego, el valor actual y máximo de los jugadores, su altura, edad, minutos jugados, y muchas otras variables importantes.

Audiencia

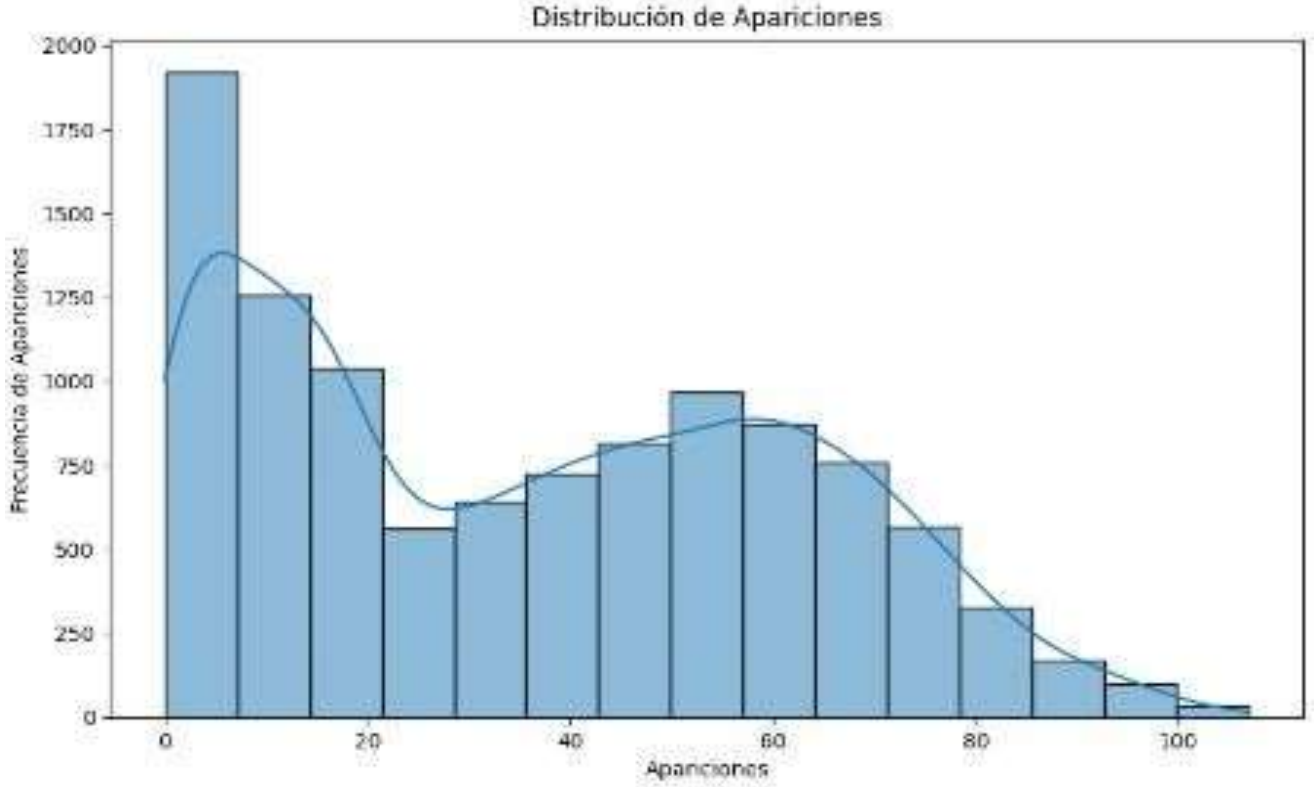
El objetivo de este informe es proporcionar una visión de negocio tanto para clubes exportadores de talento, como los equipos latinoamericanos, especialmente argentinos y brasileños, como para clubes compradores, principalmente los europeos. No obstante, este informe también puede ser útil para cualquier persona interesada, incluyendo aficionados y periodistas.

PREGUNTAS DE **INTERÉS**

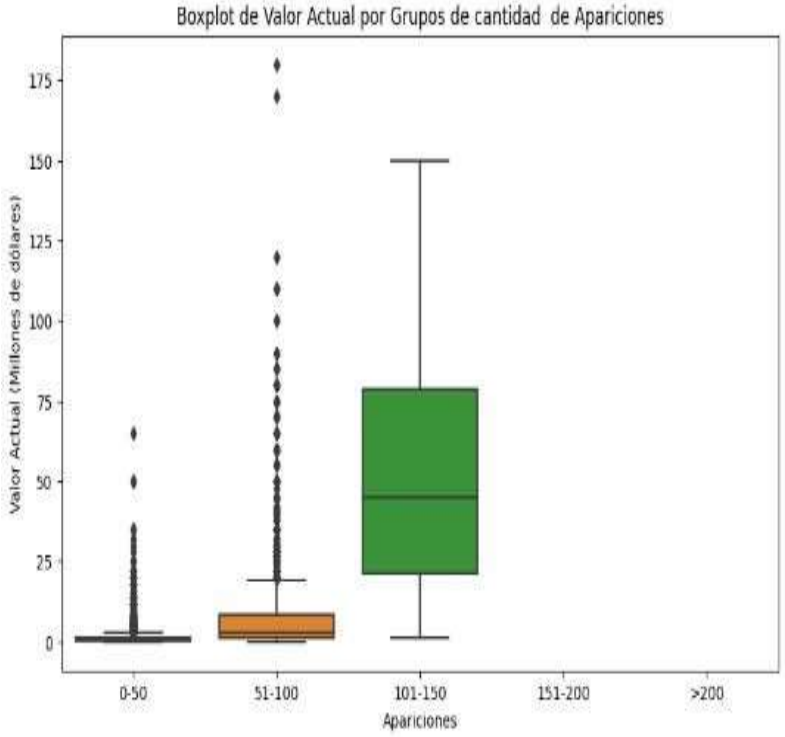
- **Preguntas principales o primarias**
- ¿Cuál es la distribución de las edades de los jugadores en el dataset?
- ¿Cuál es el valor actual más común?
- ¿Los jugadores con más apariciones tienden a tener un valor actual más alto?
- ¿Cuál es la relación entre la cantidad de apariciones y el valor actual de los jugadores?
- ¿Qué frecuencia es la que más se repite de partidos disputados?

RESUMEN **METADATA**

Frecuencia de Apariciones



Grupos por cantidad de Apariciones



ANÁLISIS **EXPLORATORIO DE DATOS (EDA)**

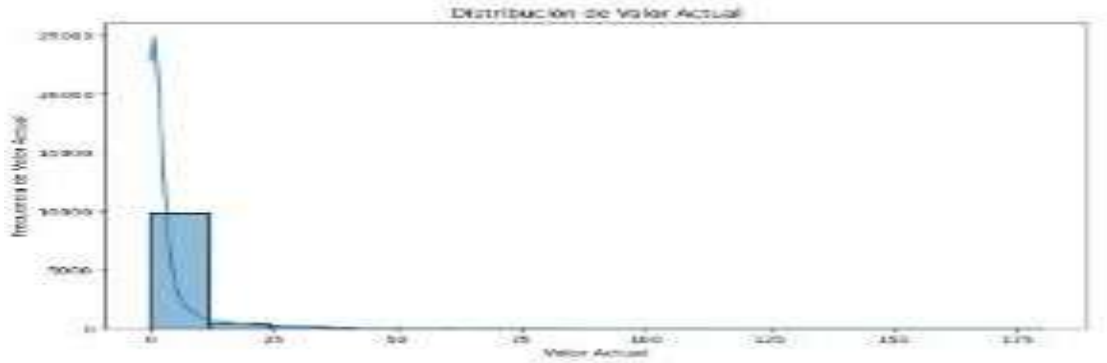
Comparación de Frecuencias

A nuestra derecha, se presentan gráficos de barras que muestran la frecuencia de:

Apariciones: La mayoría de los jugadores tienen entre 0 y 60 apariciones, con una disminución gradual en la frecuencia a medida que aumentan las apariciones. Esto sugiere una alta rotación de jugadores, con pocos acumulando un número muy alto de apariciones.

Valor Actual: La mayoría de los jugadores tienen un valor actual inferior a 10 millones de dólares. Sin embargo, hay un pequeño grupo con valores significativamente más altos, indicando la presencia de jugadores estrella en la muestra.

Frecuencia: Apariciones vs Valores Actuales

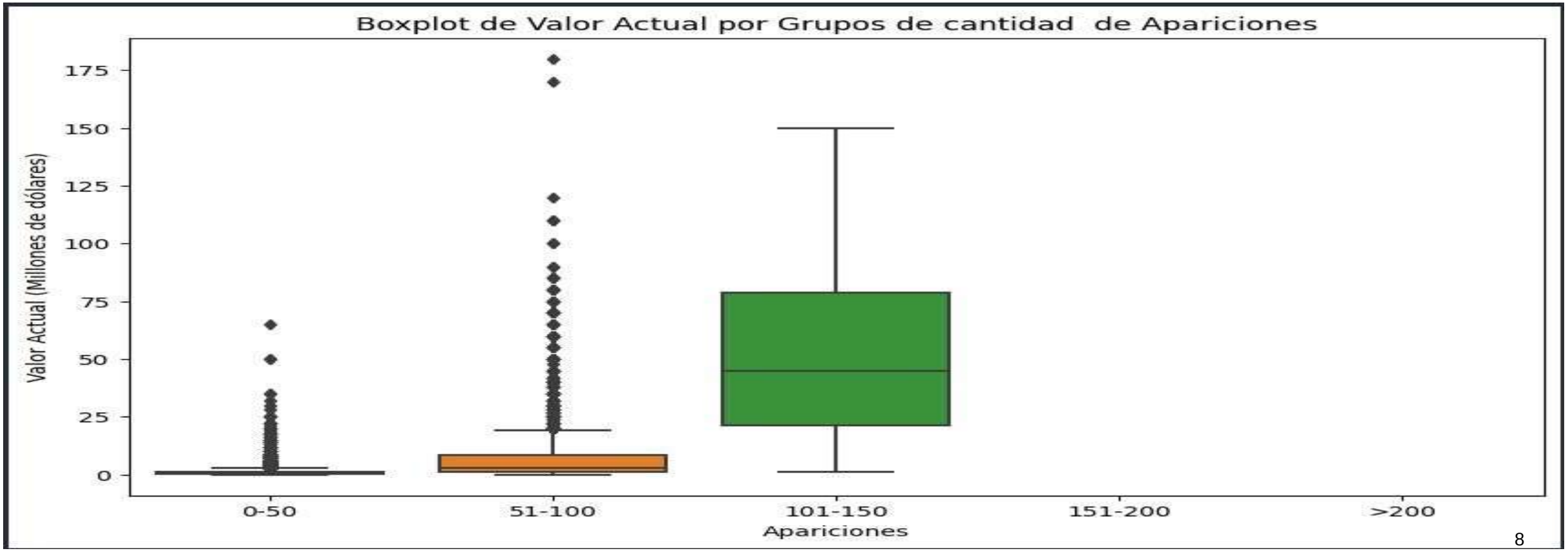


*Si bien muchas vistas pueden originarse de manera orgánica (i.e gente que hace click en un video sugerido por YouTube), vamos a asumir que la mayoría de las vistas provienen de suscriptores.

BoxPlot Relacion Valor Actual X Rangos de Apariciones.

¿Cual es el valor actual, a medida que juegan mas partidos?

En el gráfico de Boxplot, analizamos las variables dividiéndolas por rangos de apariciones. Notamos que, a medida que los jugadores juegan más partidos, la tendencia es que su valor económico aumenta. Esto sugiere que la experiencia y la exposición en el campo contribuyen a una mayor valoración de los jugadores en el mercado.



Creación de nuevas Variables.

Ingeniería de Atributos

•**Cálculo de Goles por 10 Partidos:** Se creó una nueva columna que calcula el número de goles anotados por cada jugador en un promedio de 10 partidos.
• Esto permite estandarizar el rendimiento de los jugadores en un periodo uniforme.

•**Limpieza de la Columna de Posición:** Se extrajo la posición principal de los jugadores, eliminando cualquier subcategoría o información adicional, para obtener solo la posición primaria de cada jugador.

•**Cálculo de Asistencias por 10 Partidos:** Se generó una nueva columna que normaliza las asistencias en un promedio de 10 partidos, similar al cálculo realizado para los goles.

•**Creación de la Variable de Experiencia:** Se definió una nueva métrica denominada "Experiencia", calculada como el producto de la edad del jugador y
•el número de apariciones en partidos. Esta métrica sirve como un indicador del tiempo de juego acumulado.

Modelos de Machine Learning

Creación de Modelos de Machine Learning

Regresión Lineal vs Random Forest

Se comparan dos modelos de machine learning, Regresión Lineal y Random Forest, utilizando varias métricas de evaluación para determinar cuál de ellos ofrece un mejor desempeño predictivo.

Regresión Lineal

RMSE de validación cruzada: 1.7243
MSE: 3.5462
R^2: 0.4307
MAE: 0.9948

El modelo de Regresión Lineal muestra un rendimiento modesto, con un MSE de 3.5462, lo que indica una variabilidad considerable entre las predicciones y los valores reales. El R^2 de 0.4307 sugiere que el modelo es capaz de explicar el 43.07% de la variabilidad en los datos. El MAE de 0.9948 refleja el error absoluto promedio de las predicciones.

Random Forest

RMSE de validación cruzada: 0.8682
MSE: 1.0065
R^2: 0.8384
MAE: 0.3670

El modelo de Random Forest presenta un rendimiento significativamente superior. Con un MSE de 1.0065, las predicciones son mucho más precisas, y el R^2 de 0.8384 indica que el modelo explica el 83.84% de la variabilidad en los datos. Además, el MAE de 0.3670 demuestra que las predicciones del modelo son considerablemente más exactas, con un error promedio mucho menor que el del modelo de Regresión Lineal.

Optimización del Modelo de Random Forest

Proceso de Optimización

Se realizó una búsqueda de hiperparámetros para identificar la mejor configuración posible para el modelo de Random Forest. Después de varias pruebas, se determinó que la configuración óptima incluía:

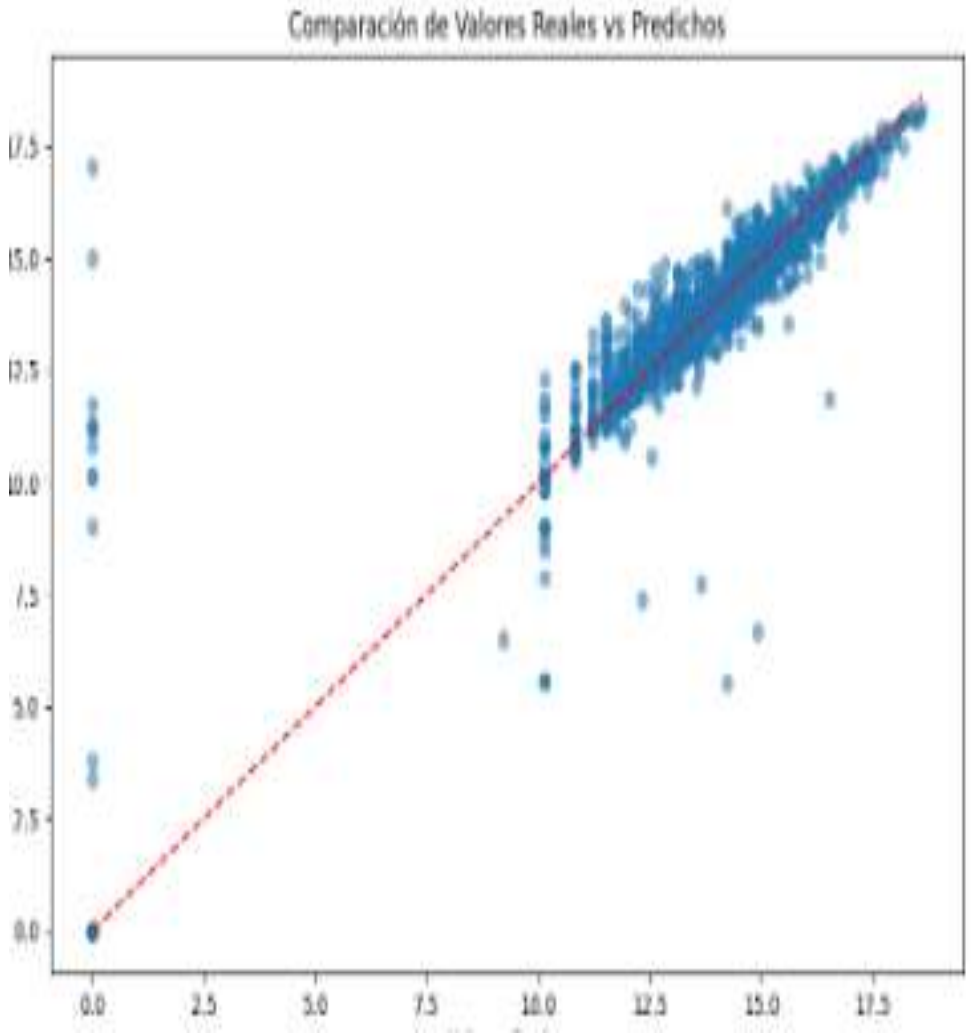
- **Número de árboles:** 100
- **Profundidad máxima:** 15
- **Mínimo de muestras por división de nodo:** 4

El modelo optimizado logró un **MSE** de 1.0175, indicando un bajo promedio de errores cuadrados en las predicciones. Además, el **R^2** obtenido fue de 0.8367, lo que sugiere que el modelo es capaz de explicar un 83.67% de la variabilidad en los datos. El **MAE** de 0.3749 revela que el error promedio absoluto es pequeño, lo que sugiere que las predicciones son precisas y consistentes.

El modelo optimizado de **Random Forest** demuestra un excelente rendimiento al predecir el valor de mercado de los jugadores. Las métricas obtenidas reflejan un buen ajuste del modelo y una baja variabilidad en los errores. La combinación de un **MSE** bajo, un **R^2** alto, y un **MAE** reducido confirma que este modelo es adecuado y confiable para la tarea, proporcionando predicciones precisas y consistentes que pueden ser utilizadas con confianza en análisis futuros.

Comparación de Valores Reales vs Predichos

Visualización de Valores Predecidos



•**Ajuste del modelo:** La mayoría de los puntos se alinean en torno a la línea roja, lo que indica que las predicciones del modelo están generalmente bien alineadas con los valores reales, lo cual es un buen signo de precisión en el modelo.

•**Dispersiones:** Algunas dispersiones visibles, especialmente para valores altos y bajos, indican que en algunos casos el modelo no predice con total exactitud,
•pero estas desviaciones no son excesivamente grandes.

•**Tendencia general:** La tendencia general de los puntos a agruparse alrededor de la línea de igualdad sugiere que el modelo realiza predicciones consistentes
• y fiables, lo que refuerza la confianza en su capacidad predictiva.