Predicción del Potencial de Jugadores FIFA 22

Análisis para toma de decisiones en fichajes y mejora de simulaciones en videojuegos.

by Gustavo Mernies

CODERHOUSE

El análisis de datos deportivos ha revolucionado la evaluación de equipos y jugadore

¿Por qué analizar datos de jugadores de FIFA 22?

1 Motivación

El análisis de datos deportivos ha revolucionado la evaluación de equipos y jugadores profesionales. Este dataset ofrece una oportunidad única para explorar la influencia de las características individuales en la valoración económica y el rendimiento percibido.

2 Insights Útiles Para:

- Clubes de fútbol (evaluación y fichajes).
- Desarrolladores de videojuegos (simulaciones realistas).

CODERHOUSE

¿A quién le interesa este análisis y por qué es importante?

Audiencia:

- Equipos de análisis deportivo: Para comprender mejor la valoración de jugadores.
- Desarrolladores de videojuegos: Para refinar las mecánicas de simulación.
- Científicos de datos y estudiantes: Como caso de estudio en predicción y análisis multivariable.

Contexto Comercial:

- Optimización de fichajes, identificación de talentos, mejora de la fidelidad en videojuegos, estrategias de marketing deportivo.

Contexto Analítico:

- Predicción de valores clave (valor de mercado, potencial), reducción de complejidad, toma de decisiones basada en datos.

¿Qué buscamos responder con este análisis?

Preguntas:

- ¿Qué características predicen mejor el potencial máximo de un jugador?
- ¿Cómo varían las valoraciones de los jugadores según su posición?

Hipótesis a Evaluar:

- ¿El valor de mercado de un jugador está correlacionado con su calificación general (‘overall’) y su potencial máximo o existen otros factores más determinantes?.

Meta principal de este proyecto

El objetivo principal es analizar y modelar los factores que influyen en el potencial de los jugadores en FIFA 22, utilizando técnicas de Machine Learning (regresión).

Esto incluye:

- Predecir el potencial máximo de un jugador a partir de sus características actuales.
- Identificar patrones clave en los atributos de los jugadores para optimizar estrategias de fichajes y scouting.
- Enfoque: Combinación de análisis descriptivo y predictivo para generar soluciones basadas en datos.

Análisis Inicial del Dataset

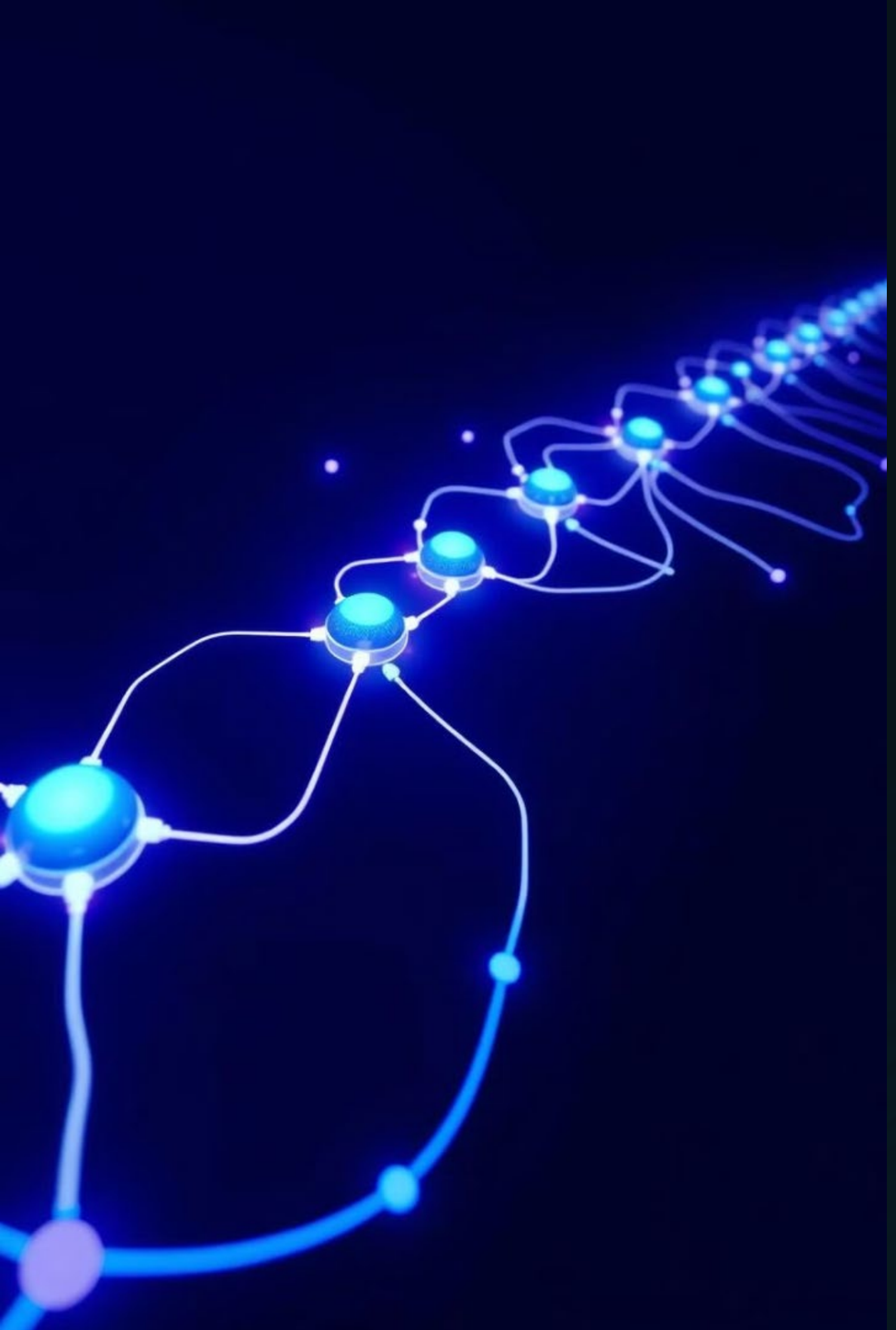
El dataset contiene 19,239 filas y 110 columnas, ofreciendo una gran cantidad de información.

Tipos de Datos: Mezcla de datetime64, float64, int64 y object.

Valores Nulos: Algunas columnas presentan valores nulos significativos, especialmente nation_logo_url (casi 96%). Sin embargo, para el análisis principal, las URL no son directamente relevantes.

Limpieza y Transformación de Datos

1. **Valores Duplicados:** No se encontraron filas duplicadas.
2. **Valores Nulos en Variables Numéricas:** Se identificaron columnas numéricas con valores nulos.
3. **Tratamiento de Valores Nulos:** Los valores nulos en value_eur y wage_eur se reemplazaron por la mediana debido a la distribución no normal de estas variables.
4. **Tratamiento de Outliers:** Se realizó un tratamiento de outliers solo para las 10 variables más correlacionadas con 'overall', reemplazando los valores extremos por la mediana en las variables económicas y por la media en las demás.



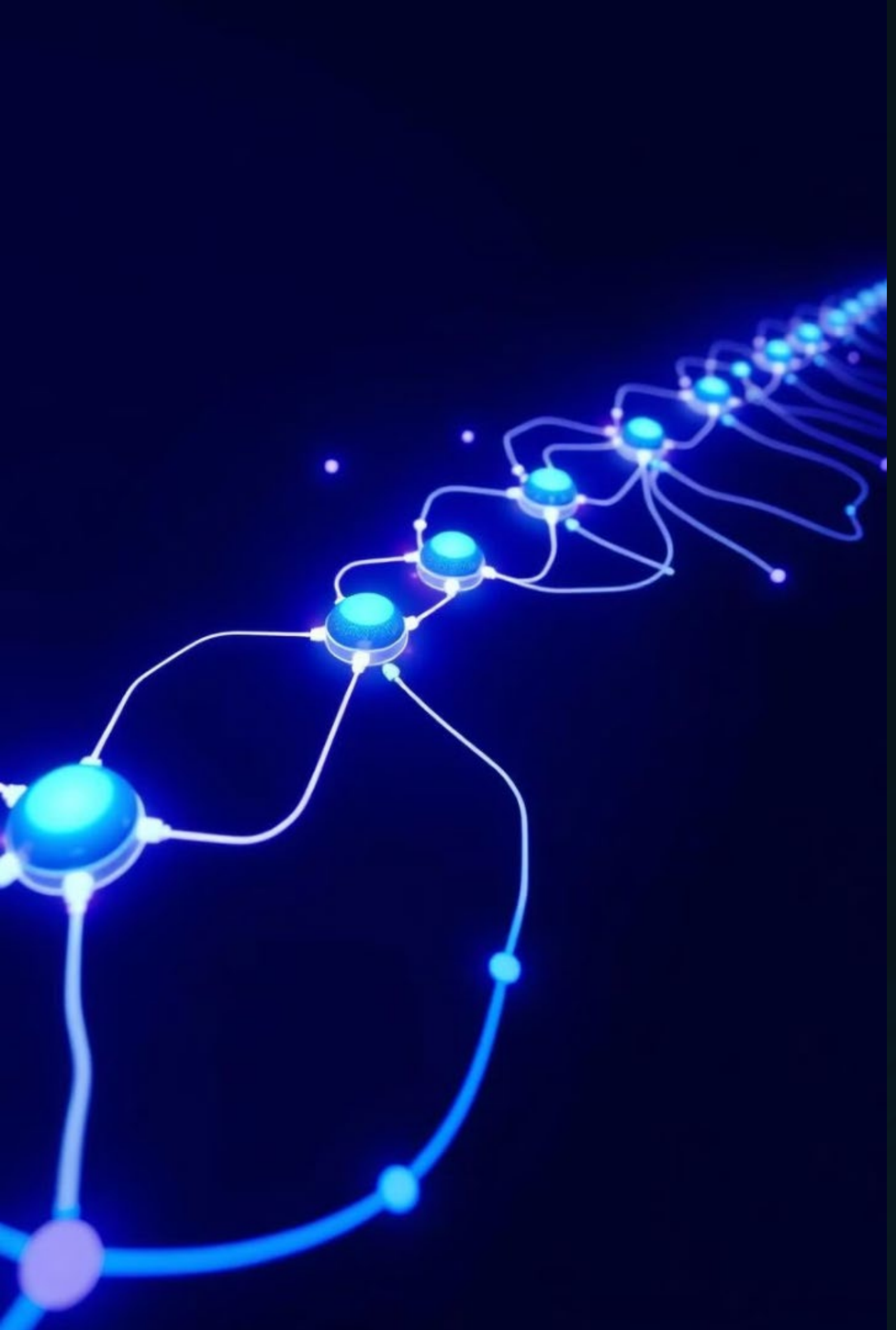
Análisis Univariado

Se analizaron las distribuciones de las 10 variables más correlacionadas con 'overall'.

Distribuciones Cercanas a la Normal: movement_reactions, passing, mentality_composure, dribbling, potential, power_shot_power, physic.

Asimetría Positiva (antes de transformación): wage_eur, value_eur, release_clause_eur mostraban colas largas y sesgo.

Transformación Logarítmica: Se aplicó una transformación logarítmica a las variables económicas para normalizar su distribución. Aunque la asimetría se redujo, persistió en cierta medida



Análisis Bivariado – Correlación con “overall”

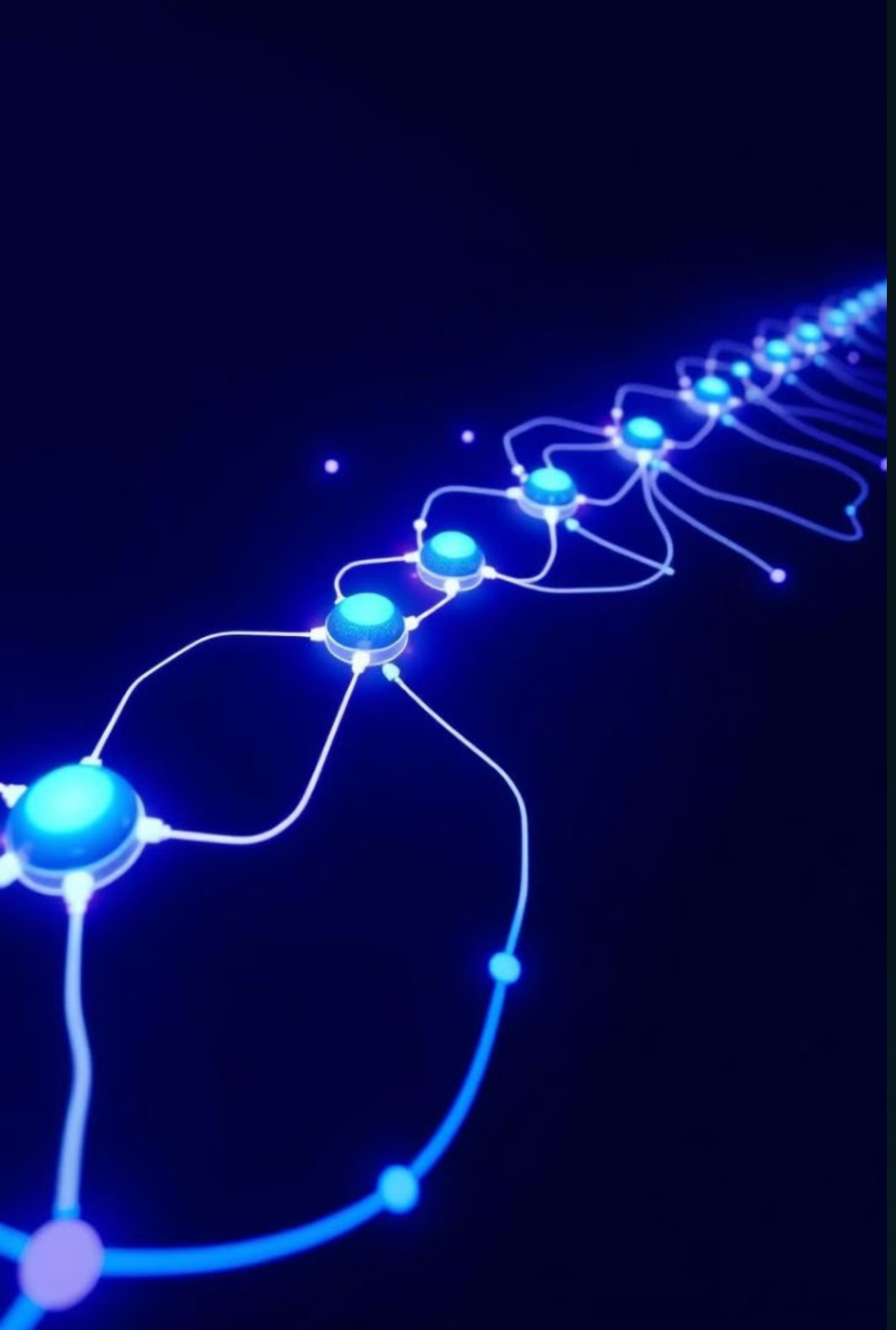
Se calcularon los coeficientes de correlación de Pearson entre las 10 variables principales y 'overall'.

Alta Correlación: value_eur (0.89), release_clause_eur (0.85), movement_reactions (0.87).

Correlación Moderada: wage_eur (0.78), dribbling (0.67), passing (0.72), mentality_composure (0.71).

Menor Correlación (pero positiva): physic (0.53), power_shot_power (0.56).

Conclusión: Tanto variables técnicas como económicas influyen en el 'overall', siendo las económicas y las reacciones las más fuertemente relacionadas.



Análisis Multivariado

Se generó una matriz de correlación para visualizar las relaciones entre las 10 variables más correlacionadas y 'overall'.

Se analizó la distribución de habilidades promedio por club, destacando a Juventus, Paris Saint-Germain y Manchester United.

Se visualizó la distribución de jugadores por nacionalidad, mostrando una alta concentración en Europa y Sudamérica (Brasil y Argentina).

•
Variables Seleccionadas: potential, value_eur, wage_eur, release_clause_eur, passing, dribbling, movement_reactions, p

Selección de Características

Se utilizó SelectKBest con la función de puntuación f_regression para seleccionar las 10 mejores variables predictoras de 'overall'.

Variables Seleccionadas:

- Potential
- value_eur
- wage_eur
- release_clause_eur
- passing, dribbling
- movement_reactions
- power_shot_power
- mentality_visión
- mentality_composure

Modelado y Evaluación

División de Datos

Conjuntos de entrenamiento y prueba (80% - 20%).

Modelos de Regresión

Regresión lineal, RandomForest y XGBoost.

Métricas de Evaluación

MSE, MAE y R^2 para valorar la precisión.

Optimización de Modelos

Se utilizó RandomizedSearchCV para optimizar los hiperparámetros de Random Forest y XGBoost.

Si bien los modelos iniciales ya tenían un buen desempeño, la optimización ayudó a encontrar la mejor configuración de parámetros.

Mejores Hiperparámetros Encontrados: (ver detalles específicos en el código).

Resultado: Se confirmó la robustez de los modelos basados en árboles para este tipo de predicción

Conclusiones Finales

Las 10 variables más correlacionadas con 'overall' fueron identificadas, incluyendo tanto atributos técnicos como económicos.

La transformación logarítmica mejoró la estabilidad de las variables económicas para el modelado.

Los modelos basados en árboles (Random Forest y XGBoost) superaron a la regresión lineal en la predicción del 'overall', alcanzando un R^2 cercano al 0.97.

Estos hallazgos son valiosos para el scouting y la toma de decisiones en el fútbol profesional..

Próximos Pasos

Profundizar en el análisis de las posiciones para ver si la correlación de ciertas variables varía según la posición.

Evaluar el impacto de la edad en el rendimiento y su relación con el salario.

Probar con otros modelos de Machine Learning y desde luego optimizar también la selección de variables e hiperparámetros.