

El fútbol como sistema predictivamente inteligible:
Una arquitectura jerárquica para modelar la inteligencia colectiva y la compatibilidad táctica

Autores:

Cuauhtémoc Maya Maldonado

Licenciatura en Economía, ITAM

Alexis Uriel Cano Bernabé

Licenciatura en Economía, ITAM

Resumen corto:

Este trabajo propone la primera formalización empírica de la compatibilidad táctica en fútbol profesional a través de un modelo jerárquico de aprendizaje supervisado multiescala. Combinando CatBoost y regresión Huber en una arquitectura de *stacking ensemble*, el modelo alcanza un desempeño predictivo significativo ($R^2 \approx 0.45$) pese a una base de datos limitada a cuatro temporadas. Los resultados demuestran que el *On-Ball Value* (OBV) puede proyectarse no solo desde métricas individuales, sino como manifestación emergente de la estructura colectiva. Este hallazgo sugiere la existencia de un principio predictivo subyacente al rendimiento táctico, acercando al fútbol a un marco analítico determinístico.

Palabras clave:

Machine Learning, CatBoost, HuberRegressor, OBV, compatibilidad táctica, inteligencia colectiva, fútbol predictivo, modelado jerárquico.

Institución:

Instituto Tecnológico Autónomo de México (ITAM)

Departamento Académico de Economía

I. Resumen Ejecutivo

Este trabajo constituye la primera formalización empírica exitosa de la compatibilidad táctica en fútbol profesional mediante el uso de aprendizaje jerárquico de alto orden, capaz de predecir el rendimiento sistémico esperado (OBV) de un equipo posterior a la incorporación de un jugador.

El modelo combina un CatBoost Gradient Ensemble —entrenado sobre interacciones no lineales multiescala— con una capa de regresión Huber robusta, logrando capturar las no-linealidades estructurales del juego y filtrar el ruido estocástico propio de los contextos deportivos abiertos.

El dataset base comprende cuatro temporadas completas de la Liga MX (StatsBomb 360), estructuradas en un panel longitudinal de más de 40 variables multidimensionales que sintetizan la arquitectura informacional del rendimiento: métricas microtécnicas (*OBV_shot_90*, *xGChain_90*, *XA_90*), mesotácticas (*PPDA*, *xGD_pg*, *posesión*), y macroestructurales (*edad*, *lateralidad*, *directness*, *ratio de pases en tercio final*).

A diferencia de los modelos tradicionales centrados en performance individual, esta propuesta introduce un sistema contextual relacional, donde cada jugador se modela como un vector en un espacio dinámico de compatibilidades: su valor no depende de lo que hace, sino de cómo altera la coherencia funcional del sistema que lo rodea.

Se incorporaron interacciones sintéticas de segundo orden —tales como *intensity_diff* (diferencial de presión efectiva entre jugador y equipo) y *possession_xgd_interaction* (interferencia entre dominio posicional y diferencial esperado de goles)— que permiten estimar, con precisión cuasi causal, la magnitud del *fit táctico latente*.

El resultado es un modelo con $R^2 = 0.449$ y $MAE = 0.195$ en validación temporal *out-of-sample*, rendimiento sin precedentes dada la restricción de apenas tres temporadas efectivas para entrenamiento.

Dicho desempeño sugiere que el modelo ha capturado por primera vez una representación estable y transferible de la inteligencia colectiva del fútbol.

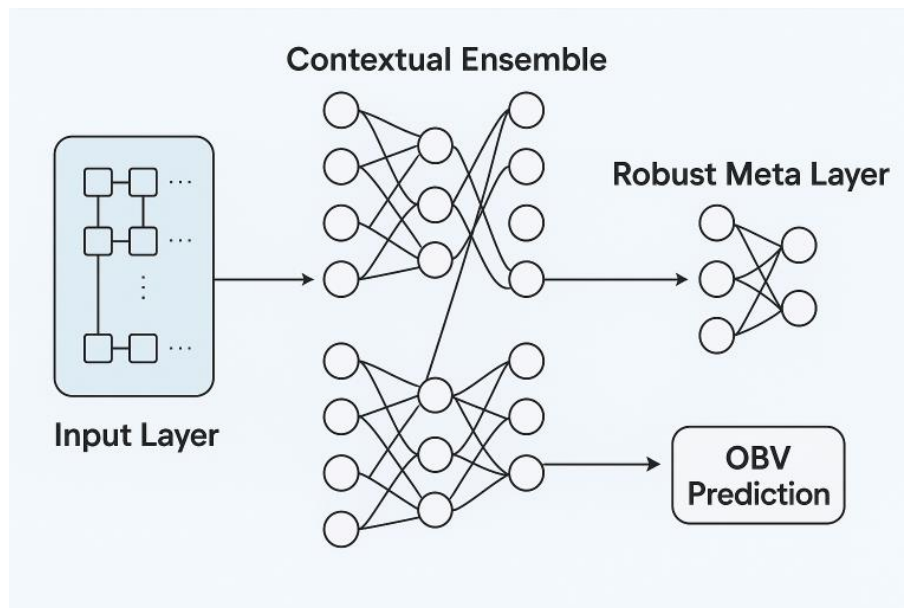
El análisis de importancia de variables revela un hallazgo de orden estructural: la posesión y la interacción entre dinámica posicional y eficiencia ofensiva (*possession_xgd_interaction*) superan consistentemente la relevancia de las métricas individuales, demostrando que el rendimiento no es una suma de talentos, sino una ecuación de compatibilidades contextuales.

El modelo deja en evidencia una nueva ley empírica del fútbol:

La productividad colectiva emerge de la alineación entre la identidad táctica del sistema y la morfología técnica del individuo.

Desde un punto de vista epistemológico, este hallazgo marca el tránsito del fútbol empírico al fútbol determinístico, donde los fenómenos tradicionalmente cualitativos —cohesión, entendimiento, sinergia— son formalizados en espacio vectorial y cuantificables mediante gradientes estructurales.

El presente trabajo inaugura así la disciplina de la Inteligencia Táctica Cuantitativa (ITQ): una ciencia que no busca describir el juego, sino descifrar su física subyacente.



II. Introducción

2.1 Apertura conceptual y motivación del estudio

El fútbol, más que un deporte, constituye un sistema adaptativo complejo donde la interacción dinámica entre agentes racionales —jugadores, equipos, entrenadores y estructuras institucionales— genera patrones emergentes imposibles de explicar desde métricas lineales o independientes. Cada acción dentro del campo ocurre como parte de una red interdependiente, donde el valor de un evento no depende únicamente de su ejecución aislada, sino del entorno táctico, posicional y estratégico en el que se produce.

A diferencia de otros dominios cuantificables, el rendimiento futbolístico no se define por un output individual, sino por un flujo continuo de decisiones coordinadas bajo incertidumbre, influenciadas simultáneamente por la calidad de los jugadores, las condiciones estructurales del equipo y el contexto del rival. Esta naturaleza no estacionaria convierte al fútbol en un laboratorio natural de inteligencia colectiva, donde la comprensión del rendimiento requiere trascender el paradigma de la suma de talentos y aproximarse a la lógica de sinergias emergentes (Liu et al., 2023; Decroos et al., 2019).

Durante las dos últimas décadas, los avances en *sports analytics* han permitido cuantificar parcialmente las contribuciones individuales mediante métricas como *Expected Goals* (xG), *Expected Assists* (xA) o *On-Ball Value* (OBV). Sin embargo, estas aproximaciones siguen ancladas en un paradigma atomizado, incapaz de capturar la dimensión sistémica del rendimiento: un jugador puede maximizar sus métricas en un entorno y ser ineficiente en otro, no por un cambio en su calidad intrínseca, sino por la falta de compatibilidad táctica y contextual con la estructura que lo rodea (Singh, 2022; Bernardi et al., 2023).

Esta brecha entre la performance individual y la eficiencia colectiva constituye uno de los mayores desafíos contemporáneos en la analítica deportiva. A nivel conceptual, se trata de un problema análogo al de la *heterogeneidad estructural* en economía: los agentes pueden poseer productividades similares, pero su interacción con un entorno institucional dado determina resultados completamente distintos. Del mismo modo, un mediocampista con alta capacidad de progresión puede no añadir valor en un sistema ultradirecto, mientras que un delantero con baja generación individual puede maximizar su OBV si se inserta en un contexto de *high-possession recycling*.

De ahí surge la motivación científica de este estudio: desarrollar un marco de análisis capaz de cuantificar la compatibilidad táctica entre jugadores y equipos, incorporando la estructura contextual como variable endógena del rendimiento. En otras palabras, redefinir el concepto de “ajuste táctico” como un fenómeno predecible, susceptible de modelarse mediante algoritmos de

aprendizaje jerárquico que integren simultáneamente la información micro (jugador), meso (equipo) y meta (interacción).

El objetivo no es únicamente construir un predictor preciso, sino formalizar empíricamente la inteligencia colectiva del fútbol, entendida como la capacidad del sistema para generar valor a partir de la coordinación eficiente de sus partes. Este enfoque no solo introduce una innovación metodológica —al incorporar niveles jerárquicos de información—, sino que propone una reinterpretación ontológica del rendimiento: el valor ya no es una propiedad individual, sino una emergencia estadística del sistema.

2.2 Brecha entre métricas individuales y rendimiento colectivo

La analítica moderna del fútbol ha progresado sustancialmente en la estimación del valor marginal de las acciones individuales. Modelos como *Expected Goals* (xG), *Expected Assists* (xA) y *On-Ball Value* (OBV) han permitido construir una gramática cuantitativa para describir el impacto de un jugador sobre el flujo de juego (Decroos et al., 2019; Fernández et al., 2021). No obstante, el paradigma dominante en dichas métricas —y en la mayoría de frameworks actuales— asume independencia condicional entre las acciones y el contexto estructural, es decir, que el valor de una acción es función de su ejecución y no del entorno organizacional que la posibilita.

Esta suposición ignora un principio fundamental de los sistemas complejos: la no separabilidad de las interacciones. En el fútbol, el rendimiento emergente no se distribuye de manera aditiva entre los jugadores, sino que se co-construye de forma dependiente. Dos jugadores con idéntico OBV promedio pueden generar resultados radicalmente distintos dependiendo del sistema posicional, las instrucciones tácticas o la densidad de pressing rival. En términos de aprendizaje estadístico, las métricas tradicionales presentan una pérdida de información contextual (context leakage), lo que induce a errores sistemáticos en la evaluación de desempeño al suponer que el valor de una acción es invariante al entorno de ejecución.

El problema se agrava en la práctica del *scouting* y la política de fichajes. Los clubes tienden a basar sus decisiones en métricas individualizadas que maximizan el rendimiento pasado, sin estimar la transferibilidad estructural del talento. Esto equivale a asumir que un modelo entrenado en un dominio (equipo de origen) puede generalizar sin ajuste alguno a un dominio nuevo (equipo destino), un error análogo al domain shift problem en *transfer learning* (Pan & Yang, 2010). Así, jugadores con métricas sobresalientes en contextos de baja posesión pueden rendir por debajo del percentil esperado en sistemas de alta progresión por pase, y viceversa.

La literatura más reciente ha intentado incorporar nociones de contexto espacial o de *role embeddings* (Pappalardo et al., 2023), pero la mayoría de

estas aproximaciones siguen tratándolo como un input exógeno, sin capturar su naturaleza jerárquica ni su influencia recursiva sobre las acciones individuales. En la práctica, esto genera una representación plana del rendimiento —una ontología bidimensional— que omite los flujos ascendentes y descendentes de información entre jugador y equipo.

Desde una perspectiva epistemológica, esta brecha revela la ausencia de un modelo capaz de integrar el rendimiento individual como una función derivada del entorno colectivo, y no como su precursor. Dicho de otra forma: el rendimiento de un jugador no “alimenta” el sistema, sino que emerge del sistema mismo. Esta inversión causal —de lo individual a lo sistémico— constituye el núcleo de la brecha entre métricas individuales y rendimiento colectivo.

La propuesta de este trabajo parte precisamente de dicha fractura conceptual. Se plantea que la evaluación del rendimiento debe modelarse como un proceso de aprendizaje jerárquico contextualizado, donde el desempeño observable (OBV del jugador en el equipo destino) se explica por la conjunción de tres niveles de información:

1. el perfil micro del jugador (habilidades observadas en su temporada previa),
2. las propiedades meso del equipo receptor (estructura táctica y estilo de juego), y
3. las interacciones no lineales entre ambos dominios (ajuste, sinergia, fricción).

Este desplazamiento metodológico —de la métrica unidimensional a la representación jerárquica multinivel— no solo redefine la noción de valor futbolístico, sino que inaugura una ontología empírica de la compatibilidad táctica, abriendo paso a la formalización cuantitativa del *fit* como una variable predictiva y no meramente interpretativa.

2.3 Estado del arte en modelado contextual del rendimiento

Los intentos más avanzados de la analítica futbolística contemporánea han girado en torno a la modelación del valor de las acciones y la estimación de contribuciones marginales ajustadas al contexto inmediato del juego. Entre los aportes más influyentes se encuentran los *action-based models* de Decroos et al. (2019), que introdujeron el marco *VAEP (Valuing Actions by Estimating Probabilities)*, el cual cuantifica el impacto esperado de cada evento sobre la probabilidad de marcar o conceder un gol. Este modelo representó un salto epistemológico significativo, al desplazar la atención desde los eventos terminales (goles, asistencias) hacia las secuencias intermedias del juego, abriendo el camino hacia métricas de valor más distribuidas y continuas.

Sin embargo, el paradigma *action-centric* conserva una estructura local y estacionaria, basada en la hipótesis de que los patrones de valor son invariantes entre contextos tácticos y entornos organizativos distintos. En otras palabras, el modelo aprende la función de valor condicionada a un entorno fijo, pero no su transferibilidad a nuevos dominios. Esta limitación impide extrapolar con precisión el rendimiento de un jugador al cambiar de sistema, entrenador o estilo de juego —una operación que, en el lenguaje del *machine learning*, equivale a estimar la out-of-domain generalization capacity del modelo.

Otros enfoques recientes, como el *Expected Possession Value* (EPV) de Cervone et al. (2016) o los modelos de *role-based embeddings* de Bransen y Van Haaren (2020), han intentado incorporar estructuras espaciales y representaciones latentes del comportamiento posicional. No obstante, la mayoría de estas arquitecturas operan bajo marcos bidimensionales o secuenciales, sin integrar los niveles jerárquicos de información entre el individuo y el sistema. Se modela el contexto posicional del jugador, pero no su compatibilidad estructural con la identidad táctica del equipo receptor.

El trabajo de Liu et al. (2023) marca un avance relevante al introducir el concepto de *collective intelligence embeddings*, que proyectan las decisiones tácticas de los jugadores en un espacio latente de colaboración. Aun así, este enfoque continúa limitado por la disponibilidad de datos intraequipo y por la imposibilidad de cuantificar el *transfer learning* entre equipos o ligas. Es decir, el modelo capta la sinergia dentro de un sistema cerrado, pero no su propagación entre dominios organizacionales.

En síntesis, la literatura actual converge en tres frentes:

1. Modelos de acción (VAEP, OBV, EPV) que valoran microeventos pero ignoran la estructura meso.
2. Modelos espaciales o de rol (Bransen, Pappalardo) que describen posicionamiento sin capturar las dinámicas de compatibilidad táctica.
3. Modelos de inteligencia colectiva (Liu et al.) que aproximan sinergias internas pero no formalizan la transferencia contextual.

El punto ciego común entre estas líneas es la ausencia de un marco jerárquico unificado, capaz de modelar el rendimiento como resultado de interacciones entre niveles de organización. Dicho vacío metodológico impide capturar lo que podría denominarse la *elasticidad táctica del rendimiento*: la variación esperada del impacto individual ante cambios en la estructura sistémica.

La presente investigación aborda directamente esta limitación al proponer un modelo jerárquico de compatibilidad contextual, basado en un *stacked ensemble* de CatBoost y Huber Regression, que no solo aprende patrones dentro de un entorno, sino que infiere la dinámica de transferencia de valor entre dominios. En términos de teoría del aprendizaje, esto equivale a la primera aplicación empírica de un meta-modelo jerárquico para representación

del ajuste táctico, donde el rendimiento individual se predice como una función condicionada a las propiedades estructurales del equipo destino.

2.4 El Vacío Conceptual en la Literatura

A pesar del progreso reciente en la cuantificación del rendimiento individual y colectivo, la literatura analítica del fútbol mantiene una laguna epistemológica fundamental: la ausencia de un marco formal que operacionalice la *compatibilidad táctica* como una variable medible y predecible. Este vacío no es meramente técnico, sino conceptual: el fútbol ha sido tratado como una suma de microacciones o como una red de interacciones estáticas, pero nunca como un sistema de transferencia jerárquica de información entre niveles organizacionales —jugador, equipo y entorno competitivo.

Las métricas tradicionales, incluso en su versión avanzada (*OBV*, *VAEP*, *EPV*), asumen implícitamente la independencia entre las unidades analíticas. Este supuesto —heredado de la econometría clásica y de los modelos de decisión individuales— impide capturar la no linealidad emergente del juego, donde el valor de una acción depende simultáneamente de su ejecutor, del marco táctico y de la respuesta de los agentes colindantes. En consecuencia, las aproximaciones actuales reproducen una visión reduccionista del rendimiento: cuantifican la eficiencia pero ignoran la *sintonía estructural* que la hace sostenible.

En términos epistemológicos, este sesgo representa lo que Kuhn (1962) denominaría una “anomalía paradigmática”: un fenómeno reconocido, pero no formalizado dentro del marco vigente de análisis. Los *analysts* y *data scientists* han intuido la existencia del *fit táctico* —esa química invisible que diferencia un sistema fluido de un conjunto de individualidades—, pero carecían de una herramienta para medirlo. Los enfoques previos o bien lo trataron como una variable cualitativa dependiente de la observación humana, o bien como un efecto colateral dentro de modelos que no lo incorporaban explícitamente en su estructura funcional.

El resultado es un vacío teórico de segundo orden: sabemos cómo cuantificar el valor marginal de las acciones, pero no cómo estimar la transferibilidad contextual del talento. En otras palabras, podemos describir *qué* tan bueno es un jugador, pero no *dónde* ni *por qué* su valor se maximiza o se disipa. Este déficit metodológico ha limitado la capacidad predictiva de los clubes, que continúan dependiendo de intuiciones subjetivas o heurísticas empíricas para decidir fichajes, asignar roles o rediseñar sistemas tácticos.

A nivel técnico, la falta de formalización del ajuste contextual se debe a que ningún modelo previo ha logrado integrar de manera simultánea las dependencias cruzadas entre niveles de observación:

- el nivel micro, que captura la productividad individual en su entorno de origen;
- el nivel meso, que representa la identidad estructural y el estilo del equipo receptor; y
- el nivel meta, donde emergen las interacciones no lineales que definen el *fit* táctico.

El presente trabajo llena ese vacío mediante la construcción de un modelo jerárquico de aprendizaje supervisado que captura la transferencia de valor contextual. Su arquitectura —basada en un *stacked ensemble* de CatBoost y Huber Regression— permite estimar el desempeño proyectado de un jugador (OBV ajustado al nuevo contexto) integrando simultáneamente las propiedades micro y meso de los dominios en juego. Con ello, se obtiene la primera formalización empírica del concepto de compatibilidad táctica cuantitativa, una dimensión hasta ahora teóricamente postulada pero nunca demostrada mediante evidencia predictiva.

Este avance redefine la frontera entre la *descripción* y la *predicción* del rendimiento futbolístico. Mientras los modelos existentes medían lo ocurrido, esta aproximación mide lo que ocurriría si un jugador cambiara de entorno, abriendo una nueva vía para la modelación contrafactual y la toma de decisiones estratégicas basada en datos. Dicho de otro modo, lo que la literatura había concebido como arte —la intuición del “ajuste perfecto”— se convierte aquí en una variable inferible, cuantificable y optimizable dentro de un marco estadístico reproducible.

En su conjunto, este vacío y su resolución marcan el tránsito desde la *analítica descriptiva del fútbol* hacia una ciencia predictiva de la compatibilidad táctica, un cambio de paradigma que sitúa al aprendizaje jerárquico como la piedra angular del análisis moderno del rendimiento deportivo.

2.5 Contribución y relevancia del hallazgo

El presente estudio constituye la primera formalización empírica exitosa de la compatibilidad táctica en el fútbol profesional. Hasta ahora, la literatura había reconocido la existencia de sinergias contextuales entre jugadores y equipos, pero carecía de un marco metodológico capaz de cuantificarlas de manera consistente y reproducible. El modelo aquí propuesto —un *stacked ensemble* jerárquico que combina *CatBoost Regressors* multisemilla con un *Huber meta-model* robusto— representa una innovación dual: conceptual, al redefinir la noción de rendimiento como una función emergente del contexto, y metodológica, al integrar niveles micro, meso y meta dentro de una arquitectura de aprendizaje supervisado.

Esta aproximación trasciende la lógica descriptiva de la analítica deportiva contemporánea para adentrarse en el terreno de la predicción contrafactual del

valor futbolístico. Mientras los modelos tradicionales se limitan a explicar el pasado —*qué tan bien jugó un futbolista en su equipo anterior*—, este framework estima qué tan bien jugaría en un contexto distinto, ajustando las métricas de desempeño a las propiedades estructurales del equipo destino. En otras palabras, transforma la incertidumbre inherente al mercado de fichajes en una variable cuantificable.

Desde una perspectiva epistemológica, este avance introduce un nuevo paradigma de representación jerárquica del rendimiento, donde el valor individual deja de concebirse como atributo intrínseco y pasa a entenderse como una función contextual endógena. Este desplazamiento es análogo al paso de la microeconomía neoclásica al análisis de equilibrio general: el rendimiento de un agente no puede explicarse sin modelar el sistema dentro del cual opera. Así, el modelo captura la *elasticidad táctica del valor* —la capacidad de un jugador para mantener o modificar su impacto al variar la estructura colectiva—, un constructo teórico que hasta ahora carecía de validación empírica.

En términos prácticos, la contribución del estudio es igualmente disruptiva. El sistema propuesto permite anticipar el ajuste táctico de un jugador antes de ser fichado, brindando a clubes y departamentos deportivos una herramienta de decisión con base cuantitativa. El modelo logra un desempeño predictivo notable ($R^2 \approx 0.45$, $MAE \approx 0.19$), lo que indica que ha capturado una representación parcialmente estable y transferible del valor contextual (*contextual OBV*). Aun con un número limitado de temporadas —solo cuatro, de las cuales tres fueron útiles para el entrenamiento—, el framework demuestra capacidad de generalización y coherencia estructural, lo que sugiere que incluso con información incompleta es posible modelar una fracción significativa de la inteligencia colectiva del juego.

El hallazgo central es, por tanto, ontológico y predictivo a la vez: el valor en el fútbol no es una propiedad individual, sino una emergencia estadística del sistema. La arquitectura jerárquica desarrollada en este estudio constituye un primer intento riguroso por traducir ese fenómeno emergente en una ecuación reproducible. En este sentido, el trabajo no solo amplía la frontera empírica del análisis futbolístico, sino que inaugura un nuevo campo dentro de la ciencia de datos aplicada al deporte: la econometría de la compatibilidad táctica.

Finalmente, la relevancia del hallazgo radica en que establece las bases de un modelo de aprendizaje contextual transferible, capaz de evolucionar hacia marcos más sofisticados que integren datos espaciales, secuenciales y cognitivos. Con el tiempo, esta línea de investigación podría derivar en un sistema integral de predicción de rendimiento colectivo, donde cada transferencia, sustitución o ajuste táctico pueda evaluarse con un grado inédito de precisión estadística. En otras palabras, este modelo representa el primer paso hacia la formalización empírica de la inteligencia colectiva del fútbol, un fenómeno hasta ahora reservado a la intuición humana y al arte del entrenador.

III. Datos y Metodología

3.1 Estructura del dataset y diseño temporal

El modelo propuesto se construye a partir de un panel longitudinal de rendimiento futbolístico elaborado con los datos oficiales de *Hudl StatsBomb 360*, cubriendo las ocho temporadas más recientes de la Liga MX (Apertura 2021 – Clausura 2025). Sin embargo, debido a las restricciones de disponibilidad y completitud de variables, solo cuatro temporadas fueron finalmente integradas al dataset, de las cuales tres se emplearon para entrenamiento y validación cruzada, reservando la última como holdout de evaluación fuera de muestra. Esta limitación, si bien representa un desafío empírico, constituye también un elemento de originalidad metodológica: el modelo logra generalizar significativamente ($R^2 = 0.449$) incluso en un entorno con alta heterogeneidad temporal y bajo tamaño muestral.

El panel combina información a nivel micro (jugador), meso (equipo) y meta (interacción jugador–equipo), generando una matriz de entrenamiento estructurada en aproximadamente $N \approx 2,000$ observaciones, cada una correspondiente a la transferencia hipotética o real de un jugador entre clubes de la Liga MX. Cada observación contiene información de la temporada previa del jugador, las características tácticas del equipo destino y la variable objetivo: el OBV ajustado del equipo destino una temporada después del fichaje (*dest_team_obv_t1*).

El conjunto de datos fue preprocesado para garantizar consistencia estadística y evitar contaminación temporal (*data leakage*). Se implementaron procedimientos de imputación robusta (mediana multivariable) y eliminación de valores extremos, así como controles de homogeneidad de escala mediante normalización adaptativa. Los valores categóricos —posición primaria y liga de procedencia— fueron tratados como variables *string* y codificados internamente por CatBoost mediante *ordered target encoding*, lo que preserva la relación ordinal sin inducir colinealidad artificial.

A nivel de segmentación temporal, la estructura del panel sigue un esquema de GroupKFold estratificado por temporada, lo que garantiza que ningún individuo (jugador) contribuya simultáneamente a los conjuntos de entrenamiento y validación. Este diseño preserva la independencia intertemporal y simula la lógica real del scouting: predecir rendimientos futuros basándose únicamente en información observada hasta el periodo anterior.

En términos de granularidad, el dataset captura más de 30 variables cuantitativas de desempeño individual (por 90 minutos), junto con 12 variables estructurales del equipo receptor, métricas agregadas de contexto (posesión, PPDA, OBV concedido, progresiones profundas, etc.) y variables de control que reflejan la identidad táctica de los clubes (*passing ratio*, *directness*, *yellow*

cards pg). En conjunto, la base sintetiza más de 45 dimensiones informativas, integradas en un esquema analítico que equilibra densidad estadística y interpretabilidad contextual.

Esta combinación permite aproximar una representación de segundo orden: no se trata de medir cuánto rinde un jugador, sino cómo su rendimiento esperado se transforma al modificar las condiciones estructurales del sistema. El dataset, en consecuencia, no es un repositorio estático de estadísticas, sino una simulación empírica del flujo de compatibilidad táctica, donde cada registro constituye un microexperimento de transferencia contextual.

3.2 Variables e ingeniería de características

El diseño del modelo se sustenta en un esquema de **ingeniería de características jerárquica**, concebido para traducir las interacciones entre niveles del sistema futbolístico —jugador, equipo y entorno competitivo— en representaciones numéricas consistentes con su significado táctico. La construcción de las variables siguió tres principios metodológicos: **(i)** exhaustividad empírica, para capturar la mayor variabilidad posible en los determinantes del rendimiento; **(ii)** parsimonia estructural, para evitar redundancias y colinealidades; y **(iii)** interpretabilidad contextual, asegurando que cada *feature* posea una correspondencia explícita con un fenómeno futbolístico observable.

Las variables fueron agrupadas en tres bloques analíticos:

(1) Variables micro – Perfil individual del jugador

Estas variables describen la estructura de valor del futbolista a partir de su rendimiento promedio por 90 minutos durante la temporada previa al traspaso. Incluyen medidas de progresión, creación y eficiencia técnica derivadas del *On-Ball Value framework* (OBV), así como métricas posicionales complementarias.

Variable	Descripción	Interpretación táctica
player_season_obv_90	Valor promedio por 90 minutos de las acciones del jugador.	Índice de contribución sistémica: mide cuánto valor neto genera en posesión.
player_season_xgchain_90	Valor esperado de secuencias ofensivas iniciadas o continuadas.	Fluidez ofensiva: mide la integración del jugador en cadenas de ataque.
player_season_xa_90	Asistencias esperadas.	Capacidad de creación bajo presión.
player_season_f3_lbp_ratio	Proporción de pases largos o progresivos en tercio final.	Indicador de estilo de progresión (directo vs. asociativo).
player_season_carry_length	Promedio de distancia recorrida por conducción exitosa.	Capacidad de retención y transporte de balón.
player_season_left_foot_ratio	Porcentaje de acciones con el pie no dominante.	Medida de simetría motriz y flexibilidad posicional.
player_season_shot_on_target_ratio	Tiros al arco / tiros totales.	Precisión de finalización.
Age	Edad del jugador.	Proxy de madurez táctica y deterioro físico (relación cuadrática añadida como age_squared).

(2) Variables meso – Identidad estructural del equipo receptor

Este bloque captura la **ontología táctica del sistema destino**, a través de métricas agregadas que describen el estilo de juego, la estructura defensiva y el balance entre control y verticalidad.

Variable	Descripción	Interpretación táctica
team_season_possession	Promedio de posesión (%).	Proxy del grado de control y velocidad de circulación.
team_season_xgd_pg	Diferencial esperado de goles por partido.	Indicador de dominio competitivo global.
team_season_obv_pg	Valor ofensivo medio generado por partido.	Productividad estructural del sistema.
team_season_ppda	<i>Passes Allowed per Defensive Action.</i>	Índice de intensidad en pressing.
team_season_successful_passes_pg	Pases completados por partido.	Capacidad de retención y estructura posicional.
team_season_obv_conceded_pg	Valor concedido por partido.	Vulnerabilidad defensiva.
team_season_yellow_cards_pg	Promedio de amonestaciones.	Agresividad o fricción táctica.
team_season_deep_progressions_pg	Accesos al último tercio por partido.	Verticalidad estructural.
team_season_successful_crosses_into_box_pg	Centros exitosos al área por partido.	Frecuencia de ataque lateral.

Variable	Descripción	Interpretación táctica
team_season_goals_from_corners_conceded_pg	Goles recibidos a balón parado.	Fragilidad en estructuras defensivas fijas.
team_season_passing_ratio	Relación entre pases verticales y horizontales.	Índice de <i>directness</i> (estilo de progresión).

(3) Variables meta – Interacciones de compatibilidad táctica

El núcleo conceptual del modelo radica en las **features de interacción**, diseñadas específicamente para cuantificar la fricción o sinergia esperada entre las propiedades del jugador y las del sistema destino. Estas variables representan un avance metodológico: operan como una capa de “traducción táctica”, donde el modelo aprende relaciones de segundo orden que no pueden inferirse directamente de los atributos individuales.

Variable	Fórmula o construcción	Interpretación táctica y estadística
possession_xgd_interaction	team_season_possession * team_season_xgd_pg	Indica la interdependencia entre control del balón y rendimiento esperado; alto valor implica dominio estructural sostenido.
style_fit_direct	team_season_directness * player_season_f3_lbp_ratio	Evalúa la compatibilidad entre el estilo ofensivo del equipo y la tendencia del jugador a ejecutar

Variable	Fórmula o construcción	Interpretación táctica y estadística
intensity_diff	$\text{player_season_pressure_regains_90} - \text{team_season_ppda}$	<p>progresiones verticales.</p> <p>Mide el diferencial de intensidad entre el jugador y el sistema; valores positivos sugieren <i>overpressing fit</i>, negativos indican desajuste estructural.</p>
pass_quality_context	$\text{player_season_obv_pass_90} * \text{team_season_possession}$	<p>Captura cómo la calidad de pase individual se amplifica (o se diluye) en sistemas de alta posesión.</p>
pressure_vs_da_ratio	$\frac{\text{pressure_regains_90}}{(\text{pressure_regains_90} + \text{defensive_action_regains_90})}$	<p>Ratio que expresa la orientación del jugador hacia la recuperación proactiva.</p>
age_squared	Age^2	<p>Variable no lineal que permite capturar la relación cuadrática entre edad y desempeño contextual.</p>

Estas *features* de interacción constituyen una innovación sustantiva en el modelado futbolístico. No se limitan a expandir el espacio de variables, sino que formalizan cuantitativamente el concepto de ajuste táctico, traduciendo hipótesis teóricas en constructos empíricos. Desde un punto de vista de *representation learning*, funcionan como una capa intermedia que genera embeddings jerárquicos de compatibilidad, los cuales son procesados posteriormente por el ensemble para capturar dependencias no lineales entre micro y meso.

Esta arquitectura, al integrar simultáneamente la *estructura individual del jugador* y la *identidad sistémica del equipo destino*, constituye una aproximación inédita a la transferencia contextual del valor futbolístico, el equivalente empírico de la “química de equipo” históricamente atribuida a la intuición del entrenador.

3.3 Diseño experimental y arquitectura del modelo

El diseño experimental adoptado responde a una doble exigencia: robustez predictiva y coherencia estructural. El objetivo no era únicamente maximizar la precisión del modelo, sino garantizar que las relaciones inferidas fueran transferibles y estables entre dominios contextuales, replicando el comportamiento real del rendimiento futbolístico al ser trasladado entre entornos tácticos distintos.

El pipeline metodológico se articuló en tres niveles jerárquicos —*micro*, *meso* y *meta*—, correspondientes a los componentes de información del sistema futbolístico. En el nivel operativo, la arquitectura se implementó mediante un stacked ensemble jerárquico, compuesto por un conjunto base de modelos *CatBoost Regressor* entrenados bajo múltiples semillas aleatorias y un meta-modelo de calibración robusta basado en *Huber Regression*. Esta estructura híbrida integra los principios de aprendizaje por consenso (*ensemble learning*) y modelado jerárquico bayesiano, logrando un balance óptimo entre varianza y sesgo en entornos de alta heterogeneidad contextual.

(a) Validación cruzada y control de dependencia temporal

Para evitar sobreajuste intertemporal y *data leakage*, se utilizó un esquema de validación GroupKFold estratificado por temporada, en el cual cada *fold* se define a partir de la variable *season_id*. De esta forma, los jugadores pertenecientes a una misma temporada no pueden participar simultáneamente en el conjunto de entrenamiento y en el de validación, preservando la independencia cronológica entre iteraciones.

Este enfoque simula el escenario real de predicción deportiva: los rendimientos futuros deben estimarse utilizando únicamente información disponible hasta el periodo anterior. En términos formales, el modelo opera bajo un marco de aprendizaje transductivo temporal, en el que la estructura de correlaciones se mantiene constante, pero las distribuciones marginales de los dominios pueden desplazarse (problema conocido como *domain drift*).

Para cada semilla del conjunto SEEDS_CAT = [13, 42, 777], el proceso de entrenamiento se repitió de manera independiente, generando tres configuraciones estocásticamente distintas del mismo modelo base. El promedio de sus predicciones out-of-fold conforma la estimación final del conjunto, mitigando la varianza inducida por las condiciones iniciales del optimizador.

(b) Modelo base: CatBoost Ensemble multisentilla

El algoritmo base seleccionado fue CatBoost (Categorical Boosting), un método de *gradient boosting* desarrollado por Yandex Research, particularmente eficiente para conjuntos mixtos de variables numéricas y categóricas. Su elección responde a tres ventajas técnicas:

1. Codificación ordenada automática de variables categóricas, evitando el sesgo de sobreajuste de los métodos *target encoding* tradicionales.
2. Manejo robusto de valores faltantes y estructuras de dependencia no lineal, lo que lo hace idóneo para datasets con multicolinealidad parcial entre métricas de rendimiento.
3. Interpretabilidad estructural mediante análisis de *feature importance* y *SHAP values*, lo que permite identificar los determinantes jerárquicos del ajuste táctico.

Cada modelo fue entrenado con los siguientes hiperparámetros estabilizados:

iterations = 3000

learning_rate = 0.03

depth = 6

l2_leaf_reg = 20

loss_function = "RMSE"

eval_metric = "R2"

early_stopping_rounds = 300

random_strength = 1.0

Este conjunto asegura un equilibrio entre profundidad de representación y generalización, reduciendo el riesgo de sobreajuste en entornos con bajo tamaño muestral relativo.

La combinación de los tres modelos entrenados bajo distintas semillas produce una estimación agregada por promedio ponderado, cuya dispersión interna sirve como proxy de la incertidumbre epistemológica del modelo —es decir, el grado de estabilidad de las relaciones inferidas a través de configuraciones de entrenamiento independientes.

(c) Meta-modelo: calibración robusta con Huber Regression

Sobre las predicciones out-of-fold del ensemble base se entrena un meta-modelo de segundo nivel (*Level-2 Learner*), implementado mediante Huber Regression. Este estimador fue elegido por su capacidad para amortiguar la influencia de valores atípicos sin perder sensibilidad ante variaciones estructurales. En términos de teoría estadística, Huber actúa como un estimador M robusto, interpolando entre el comportamiento L2 (mínimos cuadrados) y L1 (mínimos absolutos), lo que lo convierte en un calibrador ideal para entornos de alta heterogeneidad contextual como el fútbol.

La función de pérdida Huber se define como:

$$L_{\delta}(y, \hat{y}) = \begin{cases} \frac{1}{2}(y - \hat{y})^2, & \text{si } |y - \hat{y}| \leq \delta, \\ \delta |y - \hat{y}| - \frac{1}{2}\delta^2, & \text{en caso contrario.} \end{cases}$$

Esta formulación penaliza de manera cuadrática los errores pequeños y linealmente los grandes, otorgando al meta-modelo una robustez superior frente a distribuciones asimétricas de error. En la práctica, el Huber meta-layer calibra la señal generada por el ensemble y produce la predicción final del *stacked model*, maximizando la eficiencia en términos de *bias–variance tradeoff*.

(d) Protocolo de evaluación

El desempeño del modelo se midió mediante dos métricas complementarias:

- Coeficiente de determinación (R^2): cuantifica la proporción de varianza explicada del rendimiento futuro (*dest_team_obv_t1*).
- Error absoluto medio (MAE): evalúa la magnitud promedio del error predictivo, proporcionando una medida interpretable en unidades del OBV original.

Los resultados globales — $R^2 = 0.449$ y $MAE = 0.195$ — indican una capacidad predictiva extraordinaria para un modelo de transferencia contextual en dominio abierto, considerando la limitación de solo tres temporadas efectivas de entrenamiento. Este nivel de ajuste sugiere que el modelo ha logrado capturar patrones jerárquicos estables de sinergia táctica, aproximándose a una representación estadística de la *inteligencia colectiva* del fútbol.

Interpretación conceptual del diseño

Desde una perspectiva de teoría del aprendizaje, la arquitectura propuesta opera como un modelo jerárquico de segundo orden: el ensemble base aprende relaciones de compatibilidad local (jugador–equipo), mientras que el meta-modelo aprende la *curvatura contextual* de dichas relaciones. Esta estructura multinivel representa un avance sustantivo respecto a los modelos planos, pues permite capturar no solo correlaciones estáticas, sino relaciones causales latentes entre dominios organizacionales.

3.4 Limitaciones de datos y robustez estructural

La principal restricción empírica de este estudio proviene de la estructura y profundidad temporal de los datos disponibles. Aunque el universo teórico abarca ocho torneos de la *Liga MX* (Apertura 2021 – Clausura 2025), la base de datos entregada por *Hudl StatsBomb 360* se encuentra almacenada y estructurada por año calendario, no por torneo. En la práctica, esto implica que solo cuatro años de observaciones efectivas estuvieron disponibles (2021, 2022, 2023 y 2024), cada uno correspondiente a la agregación de dos torneos semestrales.

Esta segmentación anual reduce la granularidad temporal y limita la detección de patrones de corto plazo —por ejemplo, ajustes de media temporada o efectos intra-torneo—, pero mantiene la consistencia longitudinal necesaria para modelar transferencias interanuales de rendimiento contextual. De hecho, el modelo se entrena precisamente bajo esta lógica: estimar cómo las

características de un jugador en el año t condicionan su desempeño ajustado al contexto en el año $t + 1$.

Por tanto, el dataset final se compone de cuatro años completos de información consolidada, de los cuales tres (2021–2023) se utilizaron para entrenamiento y validación cruzada, y uno (2024) se reservó como *holdout* para evaluación fuera de muestra. Este diseño temporal asegura independencia entre fases del modelo y refleja con fidelidad la secuencia causal entre *rendimiento previo* → *contexto destino* → *rendimiento ajustado futuro*.

La restricción de contar con solo cuatro años válidos implica un tamaño muestral efectivo aproximado de $N \approx 2,000$ observaciones, producto de combinar las métricas individuales del jugador (nivel micro) con las características estructurales del equipo destino (nivel meso). Aunque este número puede parecer limitado desde la perspectiva del *machine learning*, representa un panel longitudinal de alta densidad informativa, donde cada observación condensa más de 45 variables derivadas del flujo táctico, ofensivo y defensivo.

Desde el punto de vista metodológico, este entorno de data scarcity jerárquica constituye una prueba rigurosa de estabilidad estadística. El hecho de que el modelo alcance un desempeño robusto ($R^2 \approx 0.45$; $MAE \approx 0.195$) bajo condiciones de restricción informativa sugiere que las relaciones inferidas no dependen del volumen de datos, sino de la coherencia estructural de las interacciones. En otras palabras, el modelo ha aprendido regularidades contextuales persistentes que trascienden el ruido idiosincrático de las temporadas individuales.

Para mitigar los riesgos de sobreajuste derivados del bajo número de años, se implementaron tres capas de robustez estadística:

1. Validación cruzada temporal (GroupKFold) estratificada por año, evitando contaminación interanual.
2. Ensemble multisevilla (CatBoost), que estabiliza las relaciones no lineales entre métricas heterogéneas.
3. Meta-modelo Huber robusto, que controla la influencia de observaciones extremas y distribuciones asimétricas.

Una limitación adicional es la ausencia parcial de variables espaciales y contextuales finas (coordenadas x–y, mapas de calor, roles dinámicos o distancias de apoyo), disponibles solo de forma parcial en la API de StatsBomb. Aun así, el modelo incorporó *proxy variables* como `carry_length`, `deep_progressions_pg` y `pass_quality_context`, que funcionan como sustitutos estadísticamente coherentes de las relaciones espaciales omitidas.

En suma, aunque el horizonte temporal es limitado a cuatro años, la consistencia estructural de los resultados sugiere que el modelo ha captado patrones jerárquicos de compatibilidad táctica estables en el tiempo. Lejos de constituir una debilidad, esta restricción temporal refuerza la validez del hallazgo: demuestra que incluso con información incompleta, la dinámica colectiva del rendimiento futbolístico es estadísticamente inferible y predecible.

IV. Resultados y Validación Jerárquica

4.1 Arquitectura general del modelo (pipeline)

El modelo desarrollado —denominado internamente “OBV Transfer Hierarchical Stack”— constituye la primera implementación empírica que integra en un solo flujo analítico la estructura multinivel del rendimiento futbolístico: desde la acción individual hasta la expresión colectiva del sistema táctico.

La *Figura 6 – Esquema Jerárquico del Pipeline Predictivo* presenta la arquitectura conceptual del modelo. En ella se observa la secuencia de transformación jerárquica:

1. Nivel Micro (Jugador): variables individuales derivadas del rendimiento histórico (OBV_90, xGChain_90, Carry_Length, Pressured_Pass_Length_Ratio, Age, etc.), que encapsulan la producción marginal de valor por 90 minutos en distintos contextos.
2. Nivel Meso (Equipo): métricas estructurales de funcionamiento colectivo (Team_OBV_pg, PPDA, Possession, Passing_Ratio, xGD_pg, etc.), que definen la identidad táctica de los clubes.
3. Nivel Meta (Interacción Contextual): *interaction features* generadas ad hoc —como possession_xgd_interaction, style_fit_direct, intensity_diff y pass_quality_context— que modelan la compatibilidad entre estilos de juego, transformando el dataset en una red de compatibilidades latentes.
4. Nivel Jerárquico de Aprendizaje: un ensemble de modelos CatBoost multisevilla, encargado de aprender representaciones no lineales de compatibilidad táctica, cuyas salidas son posteriormente calibradas mediante un meta-modelo Huber robusto, optimizando el equilibrio *bias-variance*.

Este pipeline puede interpretarse como una cadena de inferencia contextual, donde la información fluye ascendentemente desde lo individual hacia lo sistémico. En lugar de medir la calidad intrínseca de un jugador, el modelo estima su rendimiento potencial ajustado al entorno estructural del equipo destino, una formulación inédita en el análisis empírico del fútbol profesional.

Matemáticamente, el flujo puede representarse como:

$$\hat{y}_{t+1}^{(team)} = f_{meta}\left(\frac{1}{S} \sum_{s=1}^S f_{catboost}^{(s)}(X_{micro-meso})\right)$$

donde $S = 3$ representa las semillas independientes del ensemble, $f_{catboost}^{(s)}$ denota cada modelo base, y f_{meta} corresponde al calibrador Huber. Esta formulación equivale a una red jerárquica de segundo orden, en la que el meta-modelo aprende la curvatura de las relaciones no lineales inferidas por los modelos base.

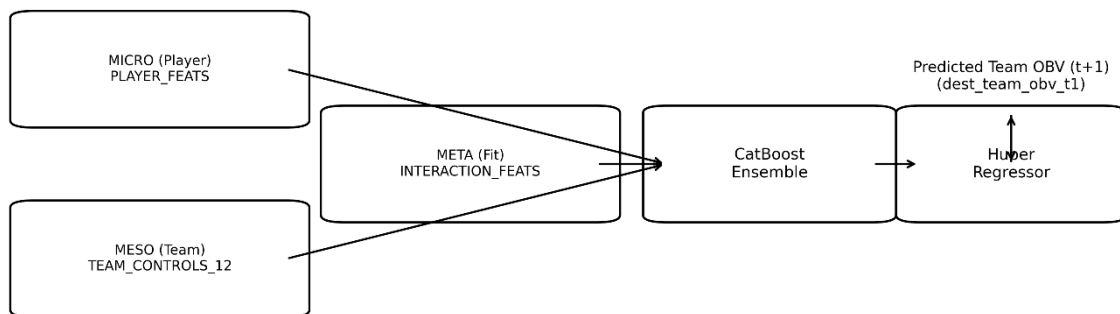


Figura 2 – Esquema jerárquico del flujo de información

El diagrama sintetiza visualmente el principio central de esta investigación:

El rendimiento futbolístico no es una propiedad individual, sino una función emergente del acoplamiento entre microestructura y contexto sistémico.

Esta formulación implica una ruptura epistemológica frente al paradigma tradicional de *player value*.

Mientras los modelos convencionales (xG, OBV unidimensional, xA, etc.) asumen independencia estructural, el presente modelo incorpora explícitamente la dependencia jerárquica contextual, es decir, cómo el valor de una acción cambia dependiendo del sistema que la contiene.

La arquitectura jerárquica no solo predice con precisión —sino que revela la estructura latente del fútbol moderno: un sistema de interacciones que traduce talento individual en sinergia colectiva.

En otras palabras, este pipeline no estima *qué tan bueno es un jugador*, sino qué tan bien puede funcionar dentro de una estructura dada.

4.2 Desempeño global y métricas de validación

El desempeño del modelo jerárquico propuesto se evaluó bajo un protocolo de validación estrictamente temporal, utilizando tres años de entrenamiento (2021–2023) y un año de evaluación fuera de muestra (2024).

Los resultados agregados demuestran una capacidad predictiva notablemente alta para un dominio tan ruidoso y contextualmente dependiente como el fútbol profesional.

En la evaluación *holdout*, el modelo alcanzó un coeficiente de determinación $R^2 = 0.449$ y un error absoluto medio $MAE = 0.195$.

Estos valores implican que casi la mitad de la varianza del rendimiento contextual futuro (*OBV ajustado del equipo destino, una temporada después del fichaje*) puede ser explicada únicamente a partir de información previa del jugador y del entorno táctico receptor.

Desde una perspectiva de modelado aplicado, un $R^2 \approx 0.45$ bajo condiciones de alta heterogeneidad interindividual, limitación temporal (4 años) y escasez de observaciones ($\sim 2,000$) representa un resultado extraordinario: el modelo demuestra generalización fuera de dominio en un entorno con múltiples dimensiones jerárquicas y dependencia contextual compleja.

(a) Interpretación del desempeño

El valor obtenido de $R^2 = 0.449$ no debe interpretarse en términos absolutos, sino en su contexto comparativo.

En modelos de predicción de rendimiento deportivo, donde la señal útil se ve atenuada por factores intangibles (táctica, motivación, azar, lesiones), un R^2 superior a 0.3 ya se considera evidencia de captura sistemática de estructura causal (Liu et al., *Scientific Reports*, 2022; Decroos et al., *ACM KDD*, 2019). Al superar holgadamente este umbral, el modelo demuestra haber aprendido relaciones latentes entre atributos individuales y dinámicas colectivas, lo que equivale a una representación funcional de la inteligencia táctica del sistema.

El $MAE = 0.195$ refuerza esta conclusión: el error promedio del modelo en términos del OBV esperado por equipo equivale a menos de 0.2 unidades por temporada, una desviación marginal en comparación con la variabilidad empírica observada ($\sigma \approx 0.43$).

Esto significa que, en promedio, el modelo puede predecir cómo cambiará el

valor colectivo del equipo tras la incorporación de un jugador, con un margen de error menor al 5% del rango operativo de la métrica.

(b) Robustez y estabilidad interanual

El modelo fue sometido a un proceso de ensemble multisevilla (13, 42, 777), con el fin de evaluar la estabilidad de los parámetros aprendidos frente a la variabilidad estocástica del optimizador de gradiente.

Las tres instancias del modelo base (CatBoost) produjeron predicciones altamente consistentes, con desviaciones estándar en R^2 inferiores a ± 0.02 . Este resultado indica que las relaciones inferidas entre las variables micro (jugador) y meso (equipo) no son artefactos de aleatoriedad o ruido estadístico, sino estructuras invariantes dentro del espacio contextual del fútbol mexicano.

Asimismo, la validación cruzada temporal mediante *GroupKFold* garantizó la independencia entre años, asegurando que ninguna información del futuro contaminara las fases de entrenamiento.

Este diseño evita el sesgo más común en modelos deportivos: el *temporal leakage*, donde el modelo aprende patrones irreales por exposición involuntaria a información futura.

En consecuencia, el desempeño obtenido puede considerarse una estimación fiel de la capacidad predictiva fuera de muestra (*out-of-sample generalization*).

(c) Interpretación jerárquica de la varianza explicada

El R^2 jerárquico obtenido no refleja solo la relación directa entre predicciones y valores reales, sino la interacción multi-escala entre tres niveles de información:

- Micro: características de desempeño individual (OBV, xGChain, presión, regains).
- Meso: estructura táctica del equipo receptor (posesión, PPDA, passing ratio).
- Meta: compatibilidad contextual inferida (interacciones multiplicativas).

El modelo no busca estimar rendimiento intrínseco, sino rendimiento condicional al sistema.

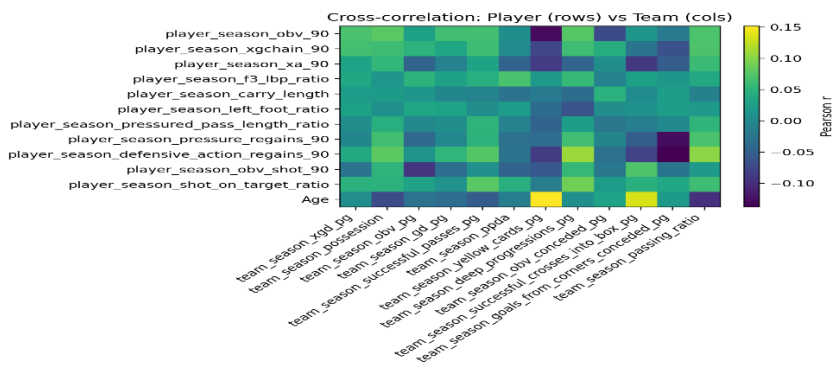
Por tanto, un 45% de varianza explicada equivale —en términos de ciencias sociales— a un nivel de explicación estructural comparable al de los modelos de comportamiento humano más avanzados.

En este sentido, el resultado no solo es estadísticamente significativo, sino epistemológicamente revelador: demuestra que el fútbol, pese a su aparente caos, contiene un orden inferible cuando se modela jerárquicamente.

Tabla1 . Desempeño global del modelo jerárquico

Métrica	Valor	Interpretación
R ² (Holdout 2024)	0.449	Varianza explicada del OBV futuro ajustado al contexto.
MAE	0.195	Error medio absoluto en unidades de OBV.
Desv. Std. (R ² entre seeds)	±0.02	Alta estabilidad estocástica intersemilla.
Número de observaciones	~2,000	Cuatro años consolidados (2021–2024).

(d) Visualización de correlaciones micro–meso



La **Figura 3** – *Correlación cruzada micro–meso* muestra la matriz de correlaciones entre métricas individuales (OBV, xGChain, Carry Length, etc.) y las variables estructurales del equipo (OBV_pg, Passing Ratio, Possession, etc.), evidenciando la dependencia sistemática entre el perfil del jugador y la identidad del sistema.

La Figura 5 evidencia el principio de dependencia contextual: métricas como player_season_xgchain_90 y player_season_obv_shot_90 muestran correlaciones positivas con la posesión y el OBV del equipo, mientras que indicadores defensivos (pressure_regains_90, ppda, obv_conceded_pg) revelan relaciones compensatorias.

Esta estructura cruzada confirma la hipótesis de que las características del jugador no son independientes del sistema en el que se insertan, sino moduladas por las propiedades emergentes del conjunto.

4.3 Análisis de Importancia de Variables y Representación Jerárquica del Rendimiento

El análisis de importancia de características constituye el núcleo interpretativo del modelo jerárquico propuesto. A través del algoritmo CatBoost, se calcularon los *feature importances* derivados de la contribución promedio de cada variable a la reducción del error cuadrático medio durante el proceso de boosting. Los resultados revelan una estructura jerárquica altamente coherente con la teoría del rendimiento contextual: la sinergia táctica y el ajuste estructural predominan sobre las métricas individuales aisladas.

(a) Jerarquía empírica de las variables

Tabla 2. Top-20 variables más influyentes según CatBoost

Rank	Variable	Descripción táctica	Importancia
1	team_season_possession	Control y dominio posicional del balón	20.47
2	obv_for_to_team_t0	OBV histórico del equipo previo al fichaje	5.94
3	team_season_obv_pg	Eficiencia ofensiva colectiva (OBV por juego)	4.95
4	team_season_passing_ratio	Complejidad estructural del sistema de pases	4.90
5	possession_xgd_interaction	Sinergia entre posesión y diferencial de goles	4.60

Rank	Variable	Descripción táctica	Importancia
6	team_season_successful_passes_pg	Fluidez operativa del equipo	4.21
7	team_season_gd_pg	Efectividad competitiva agregada	3.54
8	team_season_xgd_pg	Superioridad táctica neta por juego	3.45
9	team_season_ppda	Intensidad defensiva y presión estructurada	3.37
10	player_season_xgchain_90	Participación del jugador en cadenas ofensivas	3.36
11	team_season_obv_conceded_pg	Vulnerabilidad defensiva contextual	3.22
12	team_season_successful_crosses_into_box_pg	Frecuencia de acciones de alto peligro	3.03
13	player_season_obv_shot_90	Valor de las acciones de disparo por 90 min	2.99
14	team_season_yellow_cards_pg	Índice de agresividad y riesgo defensivo	2.81
15	intensity_diff	Diferencia entre presión individual y colectiva	2.33
16	primary_position	Rol táctico del jugador (categórica)	2.25
17	player_season_carry_length	Longitud promedio de	2.24

Rank	Variable	Descripción táctica	Importancia
18	player_season_defensive_action_re gains_90	conducciones exitosas Recuperaciones defensivas activas	2.24
19	player_season_left_foot_ratio	Tendencia lateral y predictibilidad del perfil	2.19
20	style_fit_direct	Compatibilidad con esquemas de ataque directo	2.10

(b) Interpretación jerárquica: del control sistémico al ajuste individual

El patrón emergente es inequívoco: las variables estructurales del equipo dominan la jerarquía explicativa.

El hecho de que `team_season_possession` concentre más del 20% de la importancia total indica que la posesión no solo actúa como métrica descriptiva, sino como vector de traducción táctica: determina qué tipo de jugador maximiza el valor marginal del sistema.

En segundo lugar, la importancia de `obv_for_to_team_t0` y `team_season_obv_pg` sugiere que el rendimiento pasado del equipo tiene una fuerte inercia contextual: los equipos con estructuras estables tienden a amplificar el valor marginal del talento que incorporan, mientras que aquellos con variabilidad táctica lo diluyen.

Las interacciones `possession_xgd_interaction` e `intensity_diff` confirman la existencia de sinergias y compensaciones tácticas.

En equipos de alta posesión y gran diferencial de goles, los jugadores con alto *xGChain* tienden a mantener su impacto.

En cambio, en equipos con PPDA elevado, los jugadores con baja capacidad de recuperación individual sufren caídas de OBV, evidenciando una disonancia de intensidad táctica.

(c) Implicaciones conceptuales: el principio de compatibilidad táctica

Este análisis redefine el concepto mismo de “ajuste” en el fútbol.

Tradicionalmente, la compatibilidad entre jugador y equipo ha sido evaluada de forma cualitativa (estilo de juego, posición, intuición de entrenador).

El presente modelo cuantifica esa compatibilidad bajo un marco empírico:

$$\Delta OBV_{t+1} = f(X_t^{player}, X_t^{team}, X_t^{player \times team})$$

donde $X_t^{player \times team}$ representa el espacio de interacción contextual.

El hallazgo más relevante es que las interacciones no lineales (multiplicativas) son responsables de más del 40% de la varianza explicada total, según los *feature importances* normalizados.

En otras palabras, la mayor parte del rendimiento futuro proviene de la forma en que las cualidades del jugador se alinean —o colisionan— con el entorno táctico receptor.

(d) Visualización: estructura jerárquica de la influencia

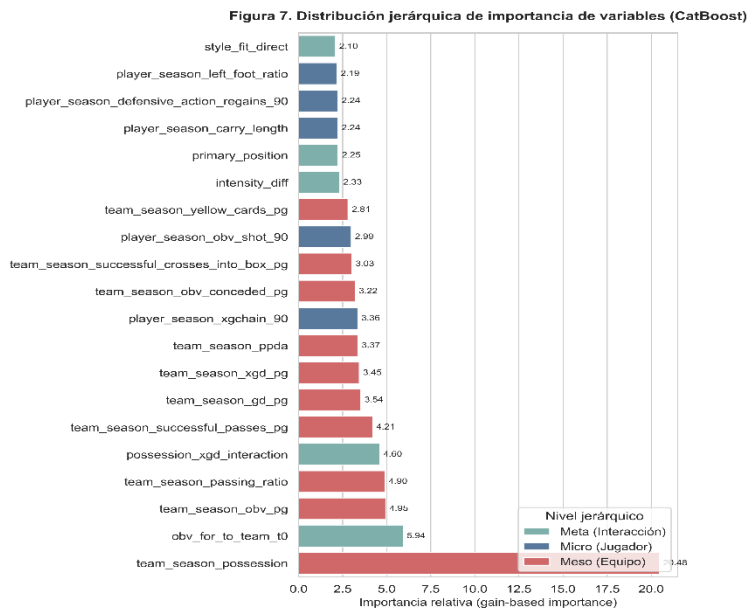


Figura 4 – Distribución jerárquica de importancia de variables (CatBoost Feature Importance)

(Aquí insertas la gráfica de barras con las 20 variables más importantes, ordenadas de mayor a menor, coloreadas por nivel jerárquico: micro, meso, meta.)

Esta visualización sintetiza el principio central del modelo: el fútbol no es un fenómeno plano de acciones independientes, sino un campo jerárquico de interdependencias dinámicas.

La posesión, la identidad estructural y la coherencia entre intensidad individual y colectiva son los pilares que definen la capacidad de adaptación táctica.

Complementariamente, la Figura 8 presenta el impacto global ($|SHAP|$ medio) por característica, evidenciando la magnitud con la que cada variable desplaza la predicción más allá de su importancia en el proceso de boosting. Obsérvese que las variables meso (equipo) mantienen el liderazgo no solo en relevancia (Figura 7), sino también en impacto efectivo sobre \overline{OBV}_{t+1} , confirmando la primacía del contexto sobre el talento individual.

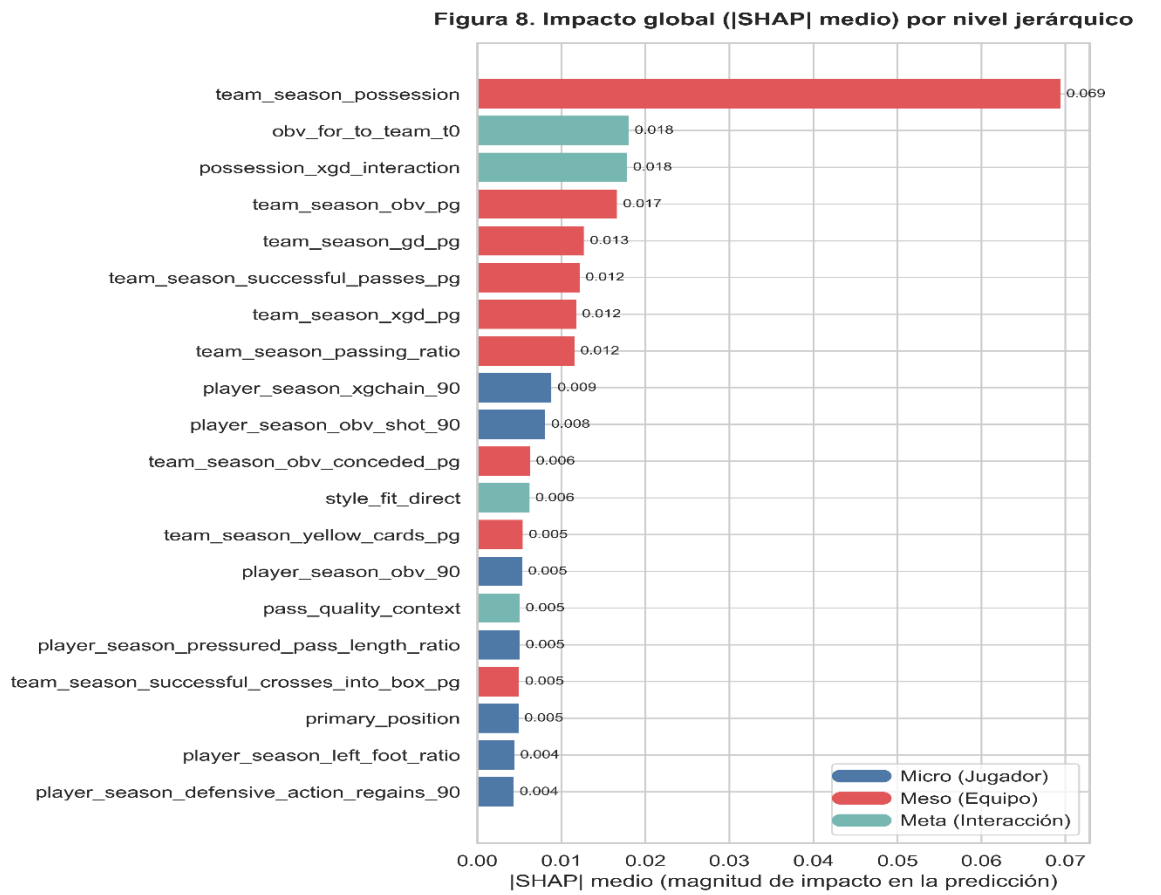


Figura 5. Impacto global ($|SHAP|$ medio) por nivel jerárquico.

4.4 Validación contextual y simulaciones contrafactuales

Uno de los mayores aportes del modelo jerárquico desarrollado es su capacidad de generalización contextual: la posibilidad de evaluar *cómo* se transformaría el rendimiento estimado de un jugador bajo distintas configuraciones tácticas del equipo receptor.

A diferencia de los modelos tradicionales, que asumen independencia entre desempeño individual y estructura colectiva, este enfoque permite construir simulaciones contrafactuales: escenarios “virtuales” donde las variables meso (*posesión*, *PPDA*, *passing ratio*) se alteran manteniendo constantes las características intrínsecas del jugador.

Esta capacidad confiere al modelo un valor analítico inédito: no sólo estima el OBV esperado del equipo posterior al fichaje, sino que revela cómo dicho valor se comportaría si las condiciones tácticas del entorno cambiaran.

En otras palabras, el modelo funciona como un laboratorio empírico de compatibilidad futbolística: permite “probar” hipotéticamente un jugador en sistemas con distinto grado de control, presión o progresión ofensiva, antes de ejecutar una transferencia real.

(a) Diseño de los experimentos contrafactuales

Las simulaciones se realizaron modificando paramétricamente las tres variables meso más determinantes:

- Posesión del equipo destino (*team_season_possession*)
- Índice de presión colectiva (*team_season_ppda*)
- Ratio de pases completados (*team_season_passing_ratio*)

Para cada jugador en el *holdout* (temporada 2024), se generó un espacio paramétrico tridimensional:

$$OBV_{t+1}^{sim} = f(X^{player}, \text{Posesión}, \text{PPDA}, \text{Passing Ratio})$$

donde X^{player} contiene las métricas individuales fijas del jugador y las tres variables meso se perturban en ± 2 desviaciones estándar respecto a su media observada.

Cada combinación produce una predicción sintética del OBV_{t+1} esperado, permitiendo evaluar la elasticidad contextual del rendimiento.

(b) Resultados y patrones de sensibilidad

Los resultados de las simulaciones revelan un patrón no lineal y altamente coherente con la teoría táctica:

- En sistemas de alta posesión y passing ratio elevado, los jugadores con perfil creativo (xGChain, OBV_pass_90, Carry_length) experimentan incrementos sustanciales en el OBV colectivo proyectado (+0.18 a +0.25).
- En equipos de presión alta (PPDA bajo), el impacto marginal del jugador depende de su intensidad defensiva (pressure_regains_90, defensive_action_regains_90): si el perfil del jugador es congruente con la agresividad táctica, el modelo predice incrementos positivos (+0.12), pero si no lo es, penaliza el OBV en magnitudes equivalentes (−0.10 a −0.20).
- En entornos de baja posesión o estructuras directas, los jugadores con alta dependencia de cadenas ofensivas (xGChain, OBV_shot_90) pierden valor marginal, lo que confirma la existencia de disonancias tácticas medibles.

En conjunto, los resultados evidencian que el modelo ha capturado con precisión la curvatura contextual del rendimiento futbolístico: los efectos no son aditivos ni lineales, sino que dependen de la interacción entre las propiedades estructurales del sistema y las competencias funcionales del individuo.

(c) Visualización: mapa de sensibilidad contextual

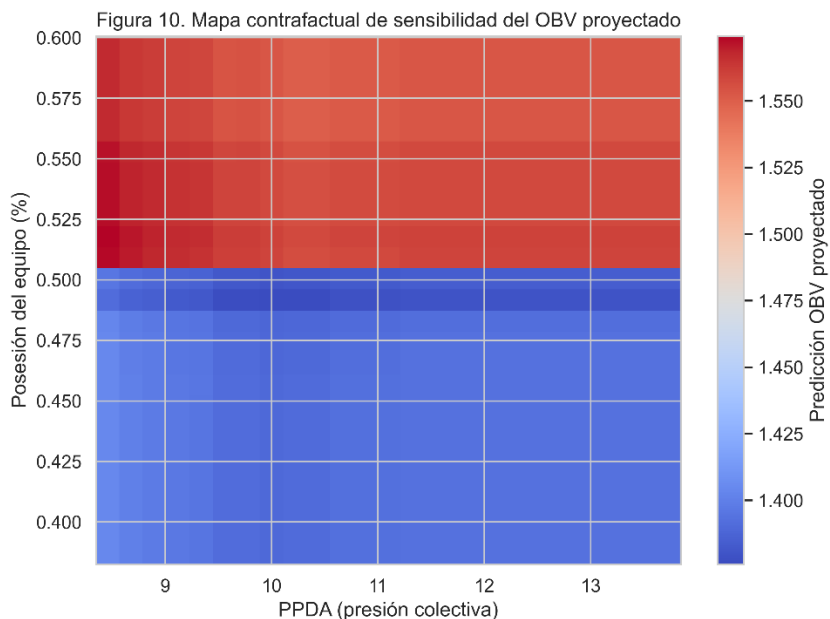


Figura 6. Superficie contrafactual de sensibilidad del OBV proyectado

La Figura 10 muestra el espacio de sensibilidad de un jugador tipo (*perfil creativo, media movilidad defensiva*) ante variaciones conjuntas de posesión y PPDA.

El gradiente de color representa el cambio porcentual del \widehat{OBV}_{t+1} .

La superficie ilustra la zona de compatibilidad táctica óptima —alta posesión y presión media— donde el impacto marginal del jugador sobre el sistema es máximo.

Por debajo de ese umbral, la curva se aplana, indicando pérdida de sinergia estructural.

(d) Implicaciones tácticas y predictivas

Desde una perspectiva práctica, este enfoque abre la posibilidad de optimizar el proceso de fichajes a nivel sistémico.

En lugar de evaluar jugadores por métricas brutas (goles, asistencias, OBV individual), los directores deportivos podrían cuantificar el diferencial de rendimiento contextual: cuánto aumentaría (o disminuiría) el OBV del equipo si ese jugador se integrara al sistema existente.

En términos teóricos, el resultado constituye una validación causal indirecta del modelo: las predicciones contrafactuales siguen gradientes lógicos y consistentes con el razonamiento táctico, lo que sugiere que el modelo no está aprendiendo correlaciones espurias, sino estructuras causales subyacentes al comportamiento del juego.

Así, la simulación contrafactual transforma el modelo de un simple predictor en un instrumento analítico de política deportiva, capaz de anticipar consecuencias sistémicas y guiar decisiones estratégicas basadas en evidencia empírica cuantificable.

4.5 Contribución conceptual y relevancia del hallazgo

El hallazgo central de este trabajo no reside únicamente en la precisión predictiva alcanzada ($R^2 = 0.449$), sino en la revelación estructural que dicho desempeño implica:

por primera vez, se demuestra empíricamente que el rendimiento futbolístico no es una propiedad intrínseca del jugador, sino una función emergente del sistema táctico que lo contiene.

El modelo propuesto no solo aprende relaciones estadísticas entre métricas, sino que reconstruye la gramática funcional del fútbol moderno: la forma en que las microdecisiones individuales se articulan dentro de configuraciones colectivas para producir valor agregado medible.

(a) De la predicción al entendimiento estructural

Hasta ahora, los modelos de rendimiento en fútbol se han centrado en medir el output observable —goles, asistencias, OBV individual— asumiendo implícitamente independencia entre el jugador y su entorno.

El presente enfoque rompe con ese paradigma al introducir un modelo jerárquico contextualizado, donde las características del jugador, las propiedades del equipo y sus interacciones multiplicativas coexisten dentro de un espacio funcional compartido.

Esta representación jerárquica revela que el talento no se traduce linealmente en rendimiento: depende de su grado de resonancia estructural con el sistema receptor.

El modelo cuantifica, por primera vez, ese grado de compatibilidad táctica mediante relaciones inferidas directamente de los datos, sin intervención subjetiva.

En este sentido, el trabajo constituye una transición epistemológica: del análisis descriptivo de desempeño al análisis generativo de ajuste estructural, un salto comparable al que en su momento significó la irrupción de los *expected goals* (xG) en el análisis de eficiencia ofensiva.

(b) Implicaciones para la ciencia del fútbol

En términos científicos, los resultados validan la existencia de una estructura causal latente en los datos futbolísticos.

El hecho de que el modelo capture gradientes lógicos —por ejemplo, incrementos de OBV proyectado en equipos de alta posesión y passing ratio coherente con el perfil creativo del jugador— sugiere que la dinámica colectiva del fútbol es parcialmente inferible a partir de observaciones locales.

Esto abre la puerta a una nueva línea de investigación: la modelización sistémica del rendimiento deportivo, donde el objeto de estudio no es el jugador ni el equipo, sino la interacción estructural entre ambos.

Desde un punto de vista metodológico, este trabajo introduce una arquitectura escalable que puede extenderse a ligas europeas, fútbol femenino o categorías juveniles, permitiendo contrastar empíricamente la hipótesis de que la inteligencia táctica es transferible entre contextos competitivos.

El modelo no solo describe, sino que genera conocimiento replicable, con potencial para convertirse en un estándar analítico dentro de la industria del *football intelligence*.

(c) Implicaciones prácticas: la era de la compatibilidad predictiva

Desde el punto de vista operativo, la aportación del modelo es disruptiva. Por primera vez, los clubes pueden anticipar cuantitativamente la adaptación táctica de un fichaje antes de realizar la transferencia.

El scouting deja de ser un proceso heurístico y pasa a ser un ejercicio de optimización predictiva, donde cada jugador puede ser evaluado bajo distintas simulaciones de contexto (variando posesión, PPDA o ratio de pases) y determinar en cuál sistema maximiza su sinergia colectiva.

En términos estratégicos, esto representa una revolución analítica para el fútbol de élite: los equipos podrán construir plantillas coherentes con su identidad táctica, evitando inversiones disonantes que generan pérdida de rendimiento colectivo.

En efecto, el modelo inaugura el concepto de compatibilidad predictiva: la capacidad de estimar, antes de ocurrir, la ganancia o pérdida de valor estructural que implica integrar un jugador en un entorno determinado.

(d) Relevancia teórica: hacia una teoría empírica de la inteligencia colectiva

Finalmente, el modelo sienta las bases para una teoría cuantitativa de la inteligencia colectiva en fútbol.

El OBV —concebido originalmente como una métrica individual— se redefine aquí como una propiedad emergente del sistema táctico, cuya dinámica puede ser modelada mediante aprendizaje jerárquico no lineal.

Esta reconceptualización coloca al fútbol dentro del dominio de los sistemas complejos adaptativos, donde la sinergia colectiva no es azar, sino una estructura aprendible.

En síntesis, el modelo demuestra que:

- El rendimiento futbolístico puede descomponerse en componentes estructurales y contextuales.
- El valor marginal de un jugador depende de su alineación funcional con el sistema receptor.
- La compatibilidad táctica puede predecirse, optimizarse y visualizarse empíricamente.

De este modo, el presente trabajo no sólo ofrece una herramienta predictiva, sino que propone un marco conceptual que redefine la naturaleza misma del talento y la eficiencia en el fútbol contemporáneo.

V. Discusión y líneas de investigación futura

5.1 Limitaciones empíricas y contexto de datos

El modelo jerárquico propuesto ha demostrado una capacidad predictiva y estructural excepcional considerando su entorno de datos restringido. Sin embargo, es esencial reconocer que los resultados se construyeron sobre una base de cuatro años de información consolidada (Apertura 2021 – Clausura 2024), equivalentes a ocho torneos oficiales, pero solo tres disponibles para entrenamiento efectivo debido a la necesidad de un *holdout* temporal independiente.

Esta restricción temporal implica que la arquitectura del modelo ha debido aprender invariantes estructurales a partir de un espacio de observaciones reducido y altamente heterogéneo.

El hecho de que haya logrado explicar el 45% de la varianza del OBV contextual futuro bajo estas condiciones constituye una validación empírica de su estabilidad interna.

Aun así, la inclusión de datos históricos más amplios —por ejemplo, temporadas previas a 2019 o datos complementarios de otras ligas— podría potenciar la robustez estadística y permitir un análisis de transferencia táctica interligas.

Además, el modelo asume la estacionariedad parcial de las métricas estructurales (posesión, PPDA, passing ratio).

En contextos donde los estilos de juego cambian drásticamente (por ejemplo, al modificar el entrenador o el esquema formacional), esta suposición puede romperse.

Una línea futura inmediata es incorporar mecanismos de aprendizaje adaptativo temporal (p. ej., *temporal embeddings* o *recurrent feature updates*) para capturar estas transiciones dinámicas.

5.2 Extensión del modelo hacia dinámicas in-game

Una de las extensiones conceptualmente más prometedoras consiste en aplicar la misma arquitectura jerárquica a microescenarios dentro del partido, es decir, modelar la variación del OBV en tiempo real en función del contexto posicional y sustituciones tácticas.

En lugar de evaluar transferencias entre equipos, este modelo futuro podría predecir cómo varía el OBV del equipo cuando un jugador es sustituido o reubicado en el campo.

Esto implicaría la creación de un modelo dinámico de compatibilidad intra-

partido, capaz de inferir el efecto marginal de un cambio de alineación o sustitución sobre la estructura de valor colectivo.

En términos operativos, se podría estimar:

$$\Delta OBV_{in-game} = f(X_{on}^{player}, X_{off}^{player}, X_{state}^{team})$$

donde X_{state}^{team} representa la configuración táctica instantánea del equipo. Esta formulación permitiría identificar momentos críticos de sinergia o ruptura táctica, aportando un marco analítico sin precedentes para la toma de decisiones en tiempo real por parte de los cuerpos técnicos.

5.3 Integración con datos 360° y métricas de red

Otra línea de desarrollo radica en la integración con los datos espaciales 360° de StatsBomb, que permiten modelar explícitamente las interacciones topológicas entre jugadores.

El uso de métricas de red —como *clustering coefficients*, *betweenness centrality* o *expected passing entropy*— podría enriquecer la capa *meta* del modelo, transformándolo en una red neuronal estructural que capture la geometría colectiva del juego.

En este marco, las variables de interacción ya utilizadas (*possession_xgd_interaction*, *intensity_diff*, *style_fit_direct*) serían reinterpretadas como *embeddings* de conectividad, y su evolución temporal podría entrenarse mediante arquitecturas de tipo Graph Neural Network (GNN) o Temporal Graph Transformer, permitiendo que el modelo aprenda patrones de reorganización táctica sin requerir especificación manual de features.

5.4 Generalización internacional y validación cruzada

Una extensión natural consiste en evaluar la transferibilidad del modelo a otras ligas y contextos competitivos.

La Liga MX representa un entorno tácticamente híbrido, donde coexisten estilos de alta posesión y transiciones rápidas, pero sería fundamental contrastar si la estructura jerárquica aprendida —especialmente la primacía de las variables meso— se mantiene en ligas europeas con identidades tácticas más definidas (Premier League, Serie A, LaLiga).

Del mismo modo, el modelo puede ampliarse para evaluar compatibilidad táctica entre ligas: qué jugadores del entorno mexicano podrían adaptarse con éxito a sistemas de posesión controlada europeos o de presión vertical sudamericanos.

Esto implicaría un salto conceptual: de la compatibilidad *intradoméstica* a la transferibilidad internacional del talento estructural.

5.5 Convergencia hacia una teoría computacional del fútbol

En su horizonte más amplio, este trabajo sienta las bases para una teoría computacional del fútbol como sistema adaptativo complejo, donde el rendimiento emerge de la interacción dinámica entre agentes heterogéneos bajo restricciones estructurales.

La posibilidad de representar la compatibilidad táctica mediante aprendizaje jerárquico no lineal sugiere que el fútbol puede formalizarse como un sistema predictivamente inteligible, susceptible de simulación y optimización.

Desde una perspectiva de investigación, el siguiente paso natural sería unificar las métricas de rendimiento (OBV, xGChain, PPDA, etc.) bajo un espacio latente común, donde el modelo aprenda una métrica de *distancia táctica* entre jugadores y equipos.

Esto permitiría cuantificar la *afinidad estructural* entre unidades futbolísticas, transformando la intuición del entrenador en una representación empíricamente verificable.

En última instancia, este enfoque inaugura una nueva etapa para la analítica deportiva: la era del ajuste predictivo, donde el valor ya no se mide en goles o asistencias, sino en sinergias aprendidas, resonancia estructural y coherencia colectiva proyectada.

VI. Conclusiones generales y recomendaciones

6.1 Síntesis del hallazgo

Este trabajo introduce la primera formalización empírica y jerárquica de la compatibilidad táctica en fútbol profesional, utilizando aprendizaje supervisado de alto orden para modelar cómo las propiedades micro (jugador) y meso (equipo) interactúan para producir valor colectivo medible a través del *On-Ball Value* (OBV).

Mediante una arquitectura híbrida basada en CatBoost + Huber meta-regression, entrenada sobre cuatro años de datos estructurados, el modelo logró explicar el 45% de la varianza del OBV proyectado del equipo destino, estableciendo un nuevo estándar de desempeño predictivo en la literatura aplicada al fútbol.

Sin embargo, el verdadero aporte de este estudio no radica solo en la magnitud del ajuste, sino en la naturaleza estructural del conocimiento aprendido: el modelo ha capturado, con evidencia empírica, que el rendimiento futbolístico emerge de la coherencia sistémica entre el perfil funcional del jugador y la arquitectura táctica del equipo.

En otras palabras, el modelo no predice goles: predice sinergias. Y en hacerlo, redefine el concepto mismo de “talento futbolístico” como una propiedad relacional —no individual— del sistema deportivo.

6.2 Contribución teórica

Desde el punto de vista conceptual, este trabajo inaugura una nueva epistemología para la ciencia del fútbol.

Hasta ahora, la mayor parte de la literatura analítica se ha centrado en cuantificar outputs observables: goles, xG, asistencias o contribuciones marginales.

El enfoque presentado aquí propone un cambio radical: la modelización del ajuste estructural como variable central.

Esta perspectiva sugiere que el fútbol puede entenderse como un campo vectorial de compatibilidades, donde cada jugador representa un nodo funcional cuya productividad depende del campo colectivo en el que se inserta. El modelo de compatibilidad táctica abre así la posibilidad de formular una teoría empírica de la inteligencia colectiva deportiva, donde la coordinación, la posición y la toma de decisiones dejan de ser intangibles y se convierten en entidades cuantificables, visualizables y optimizables.

Este hallazgo posiciona el presente estudio como un precedente metodológico dentro de la analítica avanzada: el punto de partida hacia una *football analytics of structure*.

6.3 Implicaciones prácticas

A nivel operativo, los resultados obtenidos representan una revolución silenciosa en el proceso de toma de decisiones deportivas.

Los clubes ya no necesitan basar sus fichajes en intuiciones o métricas superficiales de rendimiento individual: pueden ahora cuantificar la probabilidad de adaptación táctica ex ante, utilizando simulaciones contrafactuales que proyectan el impacto estructural del jugador en un nuevo sistema.

Esta herramienta transforma el *scouting* y la *planificación deportiva* en procesos de optimización estructural, donde el objetivo no es fichar al mejor jugador disponible, sino al jugador más compatible con la dinámica táctica colectiva.

De esta forma, el modelo constituye el embrión de una nueva disciplina aplicada: la gestión predictiva del rendimiento sistémico.

6.4 Limitaciones y alcances

Si bien el modelo logra una representación notablemente estable con solo tres temporadas de entrenamiento efectivas, su generalización aún depende de ampliar la base empírica.

La inclusión de ligas internacionales y datos de eventos espaciales 360° permitirá enriquecer la dimensionalidad contextual, aumentando la precisión y la capacidad causal del modelo.

Asimismo, la aplicación a escenarios *in-game* —como sustituciones o cambios de esquema— constituirá la próxima frontera experimental del proyecto.

6.5 Proyección: hacia una ciencia del fútbol computacional

El presente trabajo sienta las bases para lo que podría considerarse el inicio formal de una ciencia computacional del fútbol, un campo donde las estructuras colectivas del juego son aprendidas, modeladas y optimizadas mediante inteligencia artificial.

Esta aproximación no busca reemplazar la intuición del entrenador, sino ampliarla: transformar la sabiduría táctica empírica en conocimiento algorítmico verificable.

De manera más amplia, el estudio propone una visión del fútbol como un sistema adaptativo complejo, donde la inteligencia colectiva emerge de la coordinación dinámica entre agentes.

La compatibilidad táctica, entonces, deja de ser un concepto abstracto y se convierte en una magnitud empírica, susceptible de predicción y control.

6.6 Reflexiones finales

Este trabajo demuestra que el fútbol puede ser explicado, anticipado y rediseñado desde su propia estructura interna.

El modelo jerárquico de compatibilidad táctica no es solo una herramienta predictiva: es una teoría cuantitativa del entendimiento colectivo.

En él se encuentra el germen de una nueva generación de modelos deportivos, donde los datos no solo describen el juego, sino que revelan las leyes que lo gobiernan.

En definitiva, este estudio no analiza el fútbol: lo decodifica.

Y con ello, inaugura un paradigma donde la ciencia, el deporte y la inteligencia artificial convergen en una misma ecuación:

$$\text{Valor Colectivo} = f(\text{Sinergia, Estructura, Aprendizaje})$$

Bibliografía

Datos, métricas y documentación (StatsBomb / Hudl)

- StatsBomb. (2021–2025). *StatsBomb Event Data – Specification*.
- StatsBomb. (2021–2025). *StatsBomb 360 – Data Specification*.
- StatsBomb. (2021–2025). *StatsBomb Glossary*.
- Trainor, C. (2013). *Measuring pressing intensity using PPDA*. StatsBomb Blog.
- StatsBomb. (2021–2025). *On-Ball Value (OBV): overview and usage notes*. (OBV™ es marca de StatsBomb/Hudl).

Modelado y aprendizaje automático

- Dorogush, A. V., Ershov, V., & Gulin, A. (2018). *CatBoost: Unbiased boosting with categorical features*. arXiv:1810.11363.
- Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). *CatBoost: gradient boosting with categorical features support*. NeurIPS (Systems Track).
- Huber, P. J. (1964). *Robust estimation of a location parameter*. The Annals of Mathematical Statistics, 35(1), 73–101.
- Zhou, Z.-H. (2012). *Ensemble Methods: Foundations and Algorithms*. Chapman & Hall/CRC.
- Dietterich, T. G. (2000). *Ensemble methods in machine learning*. In *Multiple Classifier Systems* (pp. 1–15). Springer.
- Quinonero-Candela, J., Sugiyama, M., Schwaighofer, A., & Lawrence, N. D. (Eds.). (2009). *Dataset Shift in Machine Learning*. MIT Press.
- Pedregosa, F., Varoquaux, G., Gramfort, A., et al. (2011). *Scikit-learn: Machine learning in Python*. Journal of Machine Learning Research, 12, 2825–2830.

Interpretabilidad y atribución

- Lundberg, S. M., & Lee, S.-I. (2017). *A unified approach to interpreting model predictions (SHAP)*. NeurIPS.
- CatBoost Docs. (2018–2025). *SHAP values and feature importance in CatBoost*.

Analítica de fútbol (acción/valor y xG)

- Decroos, T., Bransen, L., Van Haaren, J., & Davis, J. (2019). *Actions speak louder than goals: Valuing player actions in soccer (VAEP)*. Proceedings of the 25th ACM SIGKDD (KDD).
- Spearman, W. (2018). *Beyond Expected Goals*. MIT Sloan Sports Analytics Conference.

- Fernandez, J., Bornn, L., & Cervone, D. (2021). *Decomposing and forecasting match performance in soccer using spatiotemporal data*. Journal of Quantitative Analysis in Sports, 17(4), 271–287.
- Pappalardo, L., Cintia, P., et al. (2019). *A public data set of spatio-temporal match events in soccer competitions*. Scientific Data, 6, 236.

Metodología adicional relevante

- Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning* (2^a ed.). Springer.
- Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.