Vedruna Vall Terrassa

Grado de especialización de desarrollo de inteligencia artificial implementado con Big Data

Análisis Futbolístico

Trabajo fin de estudio presentado por:	Mikel Zamora Torres
Tipo de trabajo:	Análisis Futbolístico
Director/a:	
Fecha:	17/01/2026

Resumen

El presente Trabajo Final de Estudios tiene como objetivo el desarrollo de un sistema basado en técnicas de *Machine Learning* para la estimación de un valor de mercado relativo de futbolistas profesionales a partir de métricas avanzadas de rendimiento. El proyecto surge de la necesidad de disponer de herramientas objetivas y reproducibles que permitan evaluar el rendimiento de los jugadores más allá de las estadísticas tradicionales, integrando además un componente visual e interactivo que facilite la interpretación de los resultados.

Para ello, se ha utilizado un conjunto de datos obtenido de la plataforma Kaggle, que contiene estadísticas avanzadas de jugadores de fútbol. El proceso metodológico ha incluido una fase de limpieza y análisis exploratorio de los datos (EDA), la generación de nuevas variables relevantes mediante *feature engineering* y el entrenamiento de distintos modelos de regresión supervisada. Concretamente, se han implementado modelos de Regresión Lineal, Ridge, Random Forest y Gradient Boosting, comparando su rendimiento para seleccionar el más adecuado.

Como resultado, los modelos basados en *ensemble*, especialmente Gradient Boosting y Random Forest, han mostrado un mejor desempeño predictivo frente a los modelos lineales. Finalmente, los resultados obtenidos se han integrado en una plataforma web desarrollada con Streamlit, complementada con visualizaciones interactivas y cuadros de mando creados en Power BI, permitiendo la comparación de jugadores y equipos de forma intuitiva.

Las conclusiones del trabajo evidencian que la combinación de *Machine Learning*, análisis de datos y desarrollo web constituye una solución válida y escalable para el análisis del rendimiento y la valoración de futbolistas en contextos reales.

Palabras clave: Machine Learning, análisis de datos, fútbol, visualización de datos, desarrollo web

Abstract

The aim of this Final Degree Project is the development of a system based on *Machine Learning* techniques to estimate a relative market value of professional football players using advanced performance metrics. The project addresses the need for objective and reproducible tools that go beyond traditional statistics, while also providing an interactive visual environment to facilitate the interpretation of the results.

A dataset obtained from Kaggle containing advanced football statistics was used. The methodology included data cleaning and exploratory data analysis (EDA), the creation of new relevant features through feature engineering, and the training of several supervised regression models. Specifically, Linear Regression, Ridge Regression, Random Forest, and Gradient Boosting models were implemented and compared in order to select the most suitable approach.

The results show that ensemble-based models, particularly Gradient Boosting and Random Forest, outperform linear models in terms of predictive accuracy. The final solution integrates the selected models into a web application developed with Streamlit, complemented by interactive dashboards built with Power BI, enabling intuitive player and team comparisons.

The conclusions demonstrate that combining Machine Learning, data analysis, and web development provides a robust and scalable solution for football performance analysis and player valuation in real-world scenarios.

Keywords: Machine Learning, data analysis, football analytics, data visualization, web development

Índice de contenidos

Índice de figuras

1. Introducción

Tradicionalmente, el valor de un futbolista se ha calculado según cosas como su fama, su edad, los equipos por los que ha pasado, su posición en el campo o cuánto se ha pagado por él en traspasos anteriores. Pero la verdad es que estos métodos no siempre muestran de manera real cómo influye un jugador en su equipo o lo que aporta realmente con su juego. Además, muchas veces estas valoraciones dependen de opiniones, lo que puede llevar a errores o decisiones poco acertadas.

Por eso, en los últimos años se ha empezado a usar Inteligencia Artificial y Machine Learning para analizar el rendimiento de los jugadores. Estas herramientas permiten trabajar con muchos datos a la vez, encontrar patrones que no se ven a simple vista y hacer predicciones más precisas que los métodos tradicionales. En nuestro caso, usamos modelos de "aprendizaje supervisado", que básicamente aprenden de datos históricos para poder estimar valores futuros.

El objetivo de este proyecto es crear un sistema que estime el valor relativo de los futbolistas de manera objetiva, usando estadísticas más avanzadas de su rendimiento. Para lograrlo, probamos varios modelos de predicción, desde uno sencillo como la regresión lineal hasta otros más complejos como KNN, Random Forest yArbol de decision. Así pudimos ver cómo cambia la precisión de las predicciones según la complejidad del modelo.

Además, hemos hecho una página web para que toda esta información sea más fácil de usar y entender. La web tiene varios apartados:

- **Scouting individual:** muestra estadísticas del jugador en su equipo actual y también jugadores similares.

- **Análisis de equipo:** permite comparar equipos, ver alineaciones y descubrir quiénes son los jugadores más cansados o con mejor rendimiento.

- **Comparativa de jugadores:** se pueden comparar estadísticas como duelos ganados, goles intentados, distancia recorrida, entre otros, y todo se puede ver con gráficos claros que hacen más fácil entender los datos.

- **Buscador de jugadores:** ayuda a los entrenadores a encontrar jugadores según minutos jugados, creación de jugadas, posición o rango de edad.

- **Ranking de jugadores:** muestra los máximos goleadores, máximos asistentes, mejores defensores y mejores medios.

En resumen, este proyecto combina análisis de datos con herramientas prácticas que pueden ayudar a entrenadores, analistas o cualquier persona que quiera conocer mejor el rendimiento de los futbolistas, y ofrece una forma más objetiva de valorar jugadores sin depender solo de opiniones o de la fama.

1.1. Motivación

La razón principal por la que hicimos este proyecto es porque nos interesaba juntar dos cosas que hoy en día son muy importantes: el fútbol profesional y la inteligencia artificial. El fútbol genera muchísimos datos sobre jugadores, equipos y partidos, pero casi nunca se usan de la mejor manera. Por eso pensamos que sería útil aplicar técnicas de análisis para sacar información que realmente sirva.

También nos motivó desde lo personal y académico, porque queríamos trabajar con modelos de Machine Learning clásicos en un problema real y complicado, que además tenga un impacto real. Estimar el valor de mercado de los jugadores es un caso perfecto, porque mezcla números, relaciones difíciles de ver y bastante incertidumbre.

Este proyecto es interesante también para la parte educativa y científica porque nos permite:

- Aplicar lo que aprendemos sobre aprendizaje supervisado con datos de verdad.
- Comparar distintos algoritmos para ver cuál funciona mejor.
- Analizar qué modelos son más fáciles de entender y cuáles predicen mejor.
- Ayudar a crear métodos que otros puedan usar para análisis deportivos.

Además, la inteligencia artificial cada vez se usa más en el fútbol. Los clubes grandes ya la utilizan para fichajes, para prevenir lesiones o para mejorar el rendimiento del equipo. Nuestro proyecto sigue esa misma línea, pero con un enfoque sencillo y basado en modelos clásicos, que cualquier analista pueda entender y usar sin ser un experto.

1.2. Planteamiento del trabajo

En el fútbol profesional, cada vez se depende más de los datos y estadísticas para tomar decisiones sobre fichajes, planificación de equipos o análisis del rendimiento de los jugadores. Sin embargo, la información disponible suele estar muy dispersa, a veces es difícil de interpretar, poco reproducible o se basa demasiado en opiniones, lo que hace que sea complicado evaluar de manera objetiva el verdadero valor de un futbolista.

El problema que detectamos es que hace falta una herramienta que sea objetiva, fácil de usar y reproducible, capaz de estimar el valor relativo de los jugadores usando sus estadísticas avanzadas y técnicas modernas de análisis de datos. Aunque el análisis de datos en deportes (Sports Analytics) está creciendo mucho, muchas de las soluciones existentes

son cerradas, difíciles de entender o requieren conocimientos técnicos avanzados, lo que limita su uso por parte de entrenadores, analistas o personas sin tanta experiencia técnica.

Por eso, este trabajo propone una solución tecnológica que combina desarrollo de software, desarrollo web y Machine Learning. La idea es transformar todos esos datos de fútbol en información útil para tomar decisiones.

La solución consiste en crear un sistema de inteligencia artificial que, usando aprendizaje supervisado, sea capaz de estimar un valor de mercado relativo de los futbolistas según sus métricas de rendimiento individual. Para ello, se prueban distintos modelos de regresión: desde modelos lineales simples, con el fin de encontrar cuál predice mejor y entender cómo se comportan los datos.

Además, desarrollamos una plataforma web interactiva que funciona como interfaz entre el usuario y el sistema. Esta página permite explorar, comparar y entender los resultados generados por los modelos de Machine Learning de manera sencilla, usando gráficos, comparativas y visualizaciones claras, sin necesidad de ser un experto en datos o programación.

Y en pocas palabras, este proyecto propone una solución completa que combina Machine Learning y desarrollo web para ofrecer una herramienta innovadora, objetiva y fácil de usar para analizar el rendimiento de los futbolistas,

1.3. Estructura del trabajo

Aquí describes brevemente lo que vas a contar en cada uno de los capítulos siguientes.

CAPÍTULO 1: Introducción

Este capítulo introduce el contexto general del trabajo y presenta los fundamentos del análisis de datos y del Machine Learning aplicados al fútbol profesional. Se expone la creciente importancia del uso de métricas avanzadas de rendimiento en el ámbito deportivo y su papel en la toma de decisiones relacionadas con la valoración de jugadores, la planificación deportiva y el análisis comparativo entre futbolistas y equipos.

Asimismo, se justifica la relevancia del proyecto dentro del contexto actual del *Sports Analytics*, destacando la necesidad de soluciones objetivas, reproducibles y accesibles que permitan transformar grandes volúmenes de datos en información útil. En este capítulo se presentan también la motivación del trabajo, el planteamiento del problema, los objetivos generales y una visión global de la contribución propuesta.

El capítulo finaliza con la descripción de la estructura de la memoria, proporcionando al lector una guía clara sobre la organización y el contenido de los capítulos siguientes.

CAPÍTULO 2: Contexto y estado del arte

Este capítulo desarrolla una revisión exhaustiva del estado del arte en el ámbito del análisis de datos aplicado al fútbol y el uso de técnicas de Machine Learning para la estimación del valor de mercado y el rendimiento de los jugadores. Se analizan trabajos académicos, informes técnicos y aplicaciones comerciales que emplean modelos estadísticos y algoritmos de aprendizaje automático para la evaluación objetiva de futbolistas.

Se revisan las principales métricas avanzadas utilizadas en el análisis del rendimiento individual, como los goles esperados (*Expected Goals*), asistencias esperadas (*Expected Assists*), acciones defensivas, progresión con balón y otras variables relevantes. Asimismo, se examinan los distintos enfoques de modelado predictivo empleados en la literatura, prestando especial atención a los modelos de regresión supervisada y a los algoritmos ensemble.

Además, el capítulo aborda el uso de herramientas de visualización de datos y plataformas web interactivas en el ámbito deportivo, destacando su importancia para facilitar la interpretación de los resultados y su accesibilidad a usuarios no técnicos. Este análisis permite contextualizar el proyecto dentro del panorama actual y justificar las decisiones técnicas adoptadas.

CAPÍTULO 3: Objetivos y metodología de trabajo

En este capítulo se definen los objetivos generales y específicos del proyecto y se detalla la metodología seguida para su desarrollo. El objetivo principal es diseñar e implementar un sistema de Machine Learning capaz de estimar un valor de mercado relativo de los futbolistas a partir de métricas avanzadas de rendimiento, así como desarrollar una plataforma web que permita visualizar y comparar dichos resultados de forma intuitiva.

Se describen las diferentes fases metodológicas del trabajo, que incluyen la recopilación y limpieza de los datos, el análisis exploratorio, la selección de variables relevantes y la preparación del conjunto de datos para el entrenamiento de los modelos. Asimismo, se detallan los modelos de aprendizaje automático utilizados, que abarcan desde modelos lineales básicos hasta algoritmos más avanzados como Random Forest.

El capítulo también explica el proceso de evaluación y comparación de los modelos, así como los criterios empleados para seleccionar el modelo final. Finalmente, se describe la metodología de desarrollo de la aplicación web, integrando los resultados del modelo predictivo con visualizaciones interactivas orientadas al usuario final.

CAPÍTULO 4: Tratamiento de datos y marco ético

Este capítulo aborda el tratamiento de los datos utilizados en el proyecto, haciendo especial hincapié en los procesos de limpieza, normalización y transformación de las variables. Se describen las técnicas empleadas para gestionar valores faltantes, detectar y corregir inconsistencias, y asegurar la calidad del conjunto de datos utilizado para el entrenamiento de los modelos de Machine Learning.

Asimismo, se analiza el marco ético y legal asociado al uso de datos deportivos. Aunque los datos empleados proceden de fuentes públicas y no contienen información personal sensible, se adoptan principios de uso responsable, transparencia y reproducibilidad. Se justifica la elección de métricas y se reflexiona sobre las limitaciones del modelo, evitando interpretaciones deterministas o conclusiones absolutas sobre el valor real de los futbolistas.

Este capítulo garantiza que el desarrollo del proyecto se ha realizado siguiendo buenas prácticas tanto desde el punto de vista técnico como ético.

CAPÍTULO 5: Desarrollo del sistema de Machine Learning

En este capítulo se presenta de forma detallada el desarrollo del sistema de inteligencia artificial propuesto. Se describen los distintos modelos de regresión implementados, comenzando por un modelo de regresión lineal como línea base (*baseline*), seguido de un modelo Ridge para introducir regularización, y continuando con modelos más avanzados como Random Forest.

Se analizan las configuraciones de cada modelo, los hiperparámetros utilizados y el proceso de entrenamiento y validación. Asimismo, se comparan los resultados obtenidos por cada algoritmo, evaluando su capacidad predictiva y su comportamiento frente a diferentes características del conjunto de datos.

Este capítulo permite comprender en profundidad las decisiones técnicas adoptadas y justifica la selección del modelo final utilizado en la aplicación web.

CAPÍTULO 6: Desarrollo de la plataforma web y visualización de resultados

Este capítulo describe el diseño y desarrollo de la plataforma web interactiva que integra el modelo de Machine Learning con herramientas de visualización de datos. La aplicación, desarrollada mediante tecnologías de desarrollo web orientadas a la ciencia de datos, permite explorar el rendimiento de los jugadores, comparar futbolistas y equipos, y analizar métricas clave a través de gráficos interactivos y paneles visuales.

Se detalla la arquitectura de la aplicación, la organización del código, el uso de componentes modulares y la integración de las funciones de análisis desarrolladas previamente. Asimismo, se explican las decisiones de diseño visual y experiencia de usuario, orientadas a facilitar la interpretación de los resultados y mejorar la accesibilidad de la herramienta.

Este capítulo demuestra la aplicación práctica de los modelos desarrollados y su transformación en un producto funcional y usable.

CAPÍTULO 7: Conclusiones y líneas de trabajo futuro

El último capítulo presenta las conclusiones del proyecto, evaluando el grado de cumplimiento de los objetivos planteados y sintetizando los principales resultados obtenidos. Se analiza la utilidad del sistema desarrollado, tanto desde el punto de vista del rendimiento del modelo de Machine Learning como de la plataforma web resultante.

Asimismo, se reflexiona sobre las limitaciones del trabajo y se proponen posibles líneas de mejora y desarrollo futuro, como la incorporación de nuevas métricas, el uso de modelos más avanzados de inteligencia artificial o la ampliación de la plataforma para incluir datos temporales y análisis predictivos más complejos.

Este capítulo finaliza destacando la contribución del proyecto al ámbito del análisis de datos deportivos y su potencial aplicación en contextos reales de toma de decisiones.

2. Contexto y estado del arte

Después de la introducción, se suele describir el contexto de aplicación. Suele ser un apartado (o dos en ciertos casos) en los que se estudia a fondo el dominio de aplicación, citando numerosas referencias. Debe aportar un buen resumen del conocimiento que ya existe en el campo de los problemas habituales identificados. Es el contexto general del trabajo.

Es conveniente que revises los estudios actuales publicados en la línea elegida, y deberás consultar diferentes fuentes. Hay que tener presente los autores de referencia en la temática del trabajo de investigación. Si se ha excluido a alguno de los relevantes hay que justificar adecuadamente su exclusión. Si por la extensión del trabajo no se puede señalar a todos los autores, habrá que justificar por qué se han elegido unos y se ha prescindido de otros.

El capítulo debería concluir con una última sección de resumen de conclusiones, resumiendo las principales averiguaciones del estudio y cómo van a afectar al desarrollo específico del trabajo.

Recuerda que debes referenciar adecuadamente los autores que citas en el texto y que en el aula virtual tienes información sobre cómo referenciar según la normativa APA.

Típicamente este capítulo se puede dividir en tres apartados:

2.1.Contexto del problema

El fútbol profesional contemporáneo se caracteriza por un alto grado de competitividad, donde la toma de decisiones estratégicas resulta determinante tanto a nivel deportivo como económico. En este contexto, la correcta evaluación del rendimiento de los futbolistas y la estimación de su valor de mercado se han convertido en aspectos críticos para clubes, analistas, agencias de representación y entidades deportivas.

Tradicionalmente, la valoración de jugadores se ha apoyado en criterios subjetivos, como la observación directa por parte de ojeadores, la experiencia de entrenadores y métricas básicas de rendimiento (goles, asistencias o minutos jugados). Si bien estos enfoques han sido ampliamente utilizados durante décadas, presentan limitaciones significativas, especialmente en un entorno donde el volumen y la complejidad de los datos disponibles han aumentado de forma exponencial.

La aparición de proveedores de datos deportivos ha facilitado el acceso a métricas avanzadas que describen el juego con un mayor nivel de detalle. Variables como los **expected goals (xG), expected assists (xAG),** acciones progresivas o métricas defensivas permiten analizar el impacto real de los futbolistas más allá de las estadísticas tradicionales. No obstante, el principal problema radica en la dificultad de integrar e interpretar este gran volumen de información de manera coherente y objetiva.

Asimismo, el valor de mercado de un jugador no depende exclusivamente de su rendimiento deportivo, sino que está influenciado por factores como la edad, la posición, el contexto competitivo y la percepción del mercado. Esta complejidad convierte la estimación del valor de mercado en un problema multidimensional, difícil de abordar mediante métodos clásicos de análisis.

En este escenario, surge la necesidad de aplicar técnicas de Machine Learning que permitan modelar relaciones complejas entre múltiples variables, así como desarrollar herramientas que faciliten la interpretación y visualización de los resultados obtenidos. La integración de modelos predictivos con plataformas web interactivas se presenta como una solución eficaz para democratizar el acceso al análisis avanzado y mejorar la toma de decisiones basada en datos.

2.2. Estado del arte

2.2.1 Análisis de datos y métricas avanzadas en el fútbol

El análisis de datos aplicado al fútbol ha experimentado un crecimiento notable en las últimas dos décadas, impulsado por la digitalización del deporte y la mejora en las tecnologías de captura de datos. Investigaciones previas han demostrado que las métricas avanzadas ofrecen una representación más precisa del rendimiento individual y colectivo, permitiendo evaluar acciones que no siempre se reflejan en los resultados finales de los partidos.

Métricas como xG y xAG se han consolidado como estándares en el análisis ofensivo, al proporcionar una estimación probabilística de la calidad de las ocasiones generadas. De forma complementaria, variables relacionadas con la progresión del balón, como pases progresivos o conducciones, permiten analizar la contribución de los jugadores a la construcción del juego. En el ámbito defensivo, las acciones combinadas de entradas e intercepciones ofrecen una medida cuantitativa de la solidez defensiva.

Diversos estudios han señalado que el uso combinado de estas métricas mejora significativamente la capacidad de análisis respecto a enfoques basados en estadísticas tradicionales. Sin embargo, también se destaca que la interpretación aislada de estas variables puede resultar engañosa, lo que refuerza la necesidad de modelos integradores.

2.2.2 Machine Learning en el análisis deportivo

El uso de técnicas de Machine Learning en el ámbito deportivo ha sido ampliamente documentado en la literatura científica. Estas técnicas permiten identificar patrones ocultos en los datos y generar modelos predictivos con un alto grado de precisión. En el fútbol, se han aplicado modelos de aprendizaje supervisado para la predicción de resultados, la detección de talento, la prevención de lesiones y la evaluación del rendimiento de jugadores.

Los modelos de regresión lineal suelen emplearse como punto de partida debido a su simplicidad y facilidad de interpretación. No obstante, numerosos estudios han evidenciado que estos modelos presentan limitaciones al enfrentarse a relaciones no lineales y a interacciones complejas entre variables. Como alternativa, se han propuesto modelos regularizados, como Ridge, que permiten mitigar problemas de sobreajuste.

Por otro lado, los algoritmos basados en árboles de decisión y métodos ensemble, como Random Forest, ha demostrado un rendimiento superior en múltiples aplicaciones deportivas. Estos modelos son capaces de capturar relaciones complejas y ofrecer estimaciones más robustas, lo que los convierte en herramientas especialmente adecuadas para problemas de predicción en entornos reales.

2.2.3 Estimación del valor de mercado de futbolistas

La estimación del valor de mercado de los futbolistas ha sido objeto de numerosos estudios, tanto en el ámbito académico como profesional. Tradicionalmente, plataformas especializadas han utilizado modelos propietarios que combinan rendimiento deportivo, edad, historial de lesiones y factores de mercado. Sin embargo, la opacidad de estos modelos limita su reproducibilidad y análisis científico.

Investigaciones recientes han explorado la posibilidad de estimar el valor de mercado a partir de métricas de rendimiento utilizando modelos de regresión supervisada. Estos trabajos destacan la relevancia de integrar variables avanzadas y aplicar técnicas de Machine Learning para mejorar la precisión de las estimaciones. Los resultados obtenidos sugieren que los modelos no lineales superan consistentemente a los enfoques lineales, especialmente cuando se dispone de un conjunto amplio de variables explicativas.

A pesar de estos avances, la mayoría de los estudios se centran exclusivamente en el desarrollo del modelo predictivo, sin abordar la necesidad de herramientas de visualización que permitan interpretar y explotar los resultados de manera práctica.

2.2.4 Plataformas web y visualización de datos deportivos

La visualización de datos constituye un elemento clave para la comunicación de resultados en el análisis deportivo. Gráficos radar, comparativas interactivas y paneles de control permiten facilitar la información compleja y facilitar visualmente la comprensión por parte de usuarios que no tiene mucha idea de futbol.

En los últimos años, han surgido diversas plataformas web orientadas al análisis futbolístico, que combinan datos avanzados con interfaces interactivas. Estas herramientas han demostrado ser especialmente útiles para la exploración de datos, el análisis comparativo entre jugadores y equipos, y la toma de decisiones estratégicas.

No obstante, muchas de estas plataformas se centran en la visualización descriptiva, sin integrar modelos predictivos personalizados ni permitir una adaptación flexible a distintos contextos de análisis. Esta limitación abre la puerta al desarrollo de soluciones que combinen Machine Learning, visualización avanzada y desarrollo web en un único sistema coherente.

2.3.Conclusiones

En conclusión, el análisis realizado pone de relieve la importancia creciente del uso de datos y técnicas de *Machine Learning* en el fútbol profesional. Las métricas avanzadas se han consolidado como una herramienta clave para evaluar el rendimiento de los jugadores, y los modelos de aprendizaje supervisado permiten afrontar retos complejos como la estimación de su valor de mercado.

No obstante, la revisión del estado del arte también muestra que existe una distancia entre el desarrollo teórico de estos modelos y su aplicación práctica. Gran parte de los trabajos se centran en mejorar la precisión de los algoritmos, dejando en segundo plano aspectos

esenciales como la interpretabilidad de los resultados y su uso mediante herramientas comprensibles y accesibles para los usuarios finales.

Ante esta situación, este proyecto plantea una solución que integra técnicas de *Machine Learning* clásico con una plataforma web interactiva. De este modo, no solo se facilita la estimación de un valor de mercado relativo de los futbolistas, sino que también se permite analizar y comparar su rendimiento de forma visual e intuitiva. Las ideas expuestas en este capítulo establecen así las bases conceptuales y técnicas sobre las que se desarrollará el proyecto en los capítulos siguientes.

3. Objetivos concretos y metodología de trabajo

Este apartado es el puente entre el estudio del dominio y la contribución a realizar. Según el tipo concreto de trabajo, el bloque se puede organizar de distintas formas, pero los siguientes elementos deberían estar presentes con mayor o menor detalle.

3.1. Objetivo general

El objetivo principal de este proyecto es crear un sistema capaz de estimar de manera objetiva el valor de los futbolistas, basándose en estadísticas reales de rendimiento y no solo en la fama, el precio de traspasos o en opiniones de expertos. Queremos demostrar que, usando inteligencia artificial y machine learning, se pueden hacer predicciones más precisas y justas sobre el aporte de un jugador a su equipo.

Lo que buscamos es ir más allá de lo que normalmente se ve en los medios: no queremos limitar el análisis a goles, asistencias o minutos jugados, sino considerar métricas más avanzadas como duelos ganados, pases claves, distancia recorrida durante los partidos, recuperaciones de balón y participación en jugadas ofensivas y defensivas. Todo esto nos permite entender **el valor real de un jugador**, cómo influye en el rendimiento colectivo y cómo se puede comparar de manera justa con otros jugadores de su posición o estilo de juego.

Además, el proyecto no se queda en la parte teórica. Queremos que los resultados sean **útiles y prácticos**: por eso desarrollamos una página web donde se pueden consultar estadísticas, comparar jugadores y ver rankings de rendimiento de manera clara y visual. Así, entrenadores, analistas o incluso aficionados pueden explorar la información de manera sencilla y tomar decisiones o formarse una opinión basada en datos reales y no solo en percepciones.

En definitiva, el objetivo general combina **análisis avanzado de datos**, **predicción mediante modelos de machine learning** y **presentación visual e interactiva**, con la idea de ofrecer una herramienta objetiva, práctica y fácil de usar que mejore la comprensión del rendimiento futbolístico.

3.2. Objetivos específicos

Para que el objetivo general del proyecto se cumpla de manera clara y práctica, nos marcamos una serie de objetivos específicos. Cada uno de ellos está pensado para cubrir un aspecto distinto del análisis, asegurando que no solo entendamos cómo valorar a un jugador, sino que también podamos hacerlo de manera útil y objetiva.

- **Analizar el rendimiento individual de los jugadores:**
  Este objetivo se centra en estudiar a los futbolistas uno por uno, pero de una manera mucho más profunda que lo que normalmente se hace en medios o en los datos clásicos. No solo queremos ver cuántos goles marca un jugador o cuántas asistencias da; nos interesa analizar cómo contribuye a su equipo en distintos niveles. Por ejemplo, podemos mirar cuántos balones recupera, cuántos duelos gana, cómo participa en jugadas ofensivas o defensivas, y hasta cuánto corre durante un partido. Todo esto nos da una visión mucho más completa de su influencia dentro del campo. La idea es que, usando estos datos, podamos comparar su rendimiento real con lo que la gente podría pensar a simple vista. Esto también sirve para detectar jugadores que quizá no destacan en goles, pero que son fundamentales para el equipo por otras métricas.

- **Comparar jugadores entre sí:**
  Uno de los objetivos más interesantes es poder comparar a distintos jugadores de manera objetiva. En el fútbol, muchas veces las comparaciones se hacen según la fama o el club en el que juega un futbolista, pero eso no siempre refleja lo que hace en el campo. Con nuestro sistema, podemos comparar jugadores que juegan en distintas ligas o equipos, mirando estadísticas avanzadas. Por ejemplo, podemos ver que un centrocampista joven en un equipo modesto recupera más balones y tiene más pases clave que otro en un equipo grande, aunque este último tenga más minutos de juego o más títulos. Esto permite identificar talento oculto y jugadores con potencial que aún no han recibido la atención que merecen. Además, ayuda a entrenadores y analistas a tomar decisiones más informadas sobre fichajes o

alineaciones.

- **Evaluar el rendimiento de los equipos:**

  Este objetivo busca mirar más allá del jugador individual y analizar cómo funciona un equipo en conjunto. No todos los equipos rinden igual aunque tengan jugadores estrella; a veces, la combinación de ciertos jugadores, su estado físico y su sincronización determina el rendimiento real. Por ejemplo, nuestro sistema permite identificar qué jugadores están más cansados, cuáles destacan en ciertas posiciones o cómo afectan los cambios de alineación al rendimiento global. También podemos ver patrones de juego: equipos que mantienen la posesión, que presionan alto o que dependen de un solo goleador. Esto es súper útil para entrenadores que quieren ajustar tácticas o planificar descansos estratégicos para ciertos jugadores.

- **Desarrollar una herramienta visual y accesible:**

  Otro objetivo clave es que toda esta información no se quede solo en un conjunto de datos o cálculos complejos, sino que sea **fácil de usar y entender**. Por eso hemos creado una página web donde se pueden consultar estadísticas de jugadores, rankings, comparaciones y análisis de equipos de manera visual. Los gráficos interactivos, las tablas y los filtros permiten explorar los datos sin necesidad de ser un experto en estadísticas o programación. Por ejemplo, un entrenador puede filtrar jugadores por posición y minutos jugados, y ver al instante quién tiene mejores métricas de pases clave o duelos ganados. Esto hace que la información sea **útil y aplicable**, no solo un montón de números.

- **Validar y ajustar los modelos de predicción:**

  Finalmente, queremos asegurarnos de que todo lo que haga el sistema sea fiable. Por eso uno de nuestros objetivos es validar los modelos de machine learning que usamos, ajustando parámetros y probando los resultados con datos reales de temporadas pasadas y actuales. Esto nos permite comprobar que las predicciones no son aleatorias y que reflejan de manera realista el valor y rendimiento de los jugadores. Además, si detectamos errores o inconsistencias, podemos corregirlos y mejorar continuamente el sistema. Este objetivo garantiza que nuestra herramienta

sea **precisa, confiable y útil** para cualquiera que quiera analizar fútbol de manera objetiva.

3.3. Metodología del trabajo

La metodología de este proyecto combina **análisis de datos**, **machine learning** y **desarrollo web**, siguiendo un enfoque práctico e iterativo, parecido a lo que se haría en CRISP-DM, pero adaptado a nuestras necesidades y estilo de trabajo como estudiantes. Lo dividimos en varias fases:

- **Recopilación de datos:**
  El primer paso fue buscar y descargar estadísticas de jugadores y equipos. Nos centramos tanto en datos básicos (goles, asistencias, minutos jugados) como en datos avanzados, que son los que realmente permiten medir la influencia de un jugador en el equipo, como duelos ganados, precisión de pases, balones recuperados y distancia recorrida. Esto implicó combinar información de diferentes fuentes y asegurarnos de que los datos fueran consistentes y fiables.

- **Limpieza y preparación de datos:**
  Una vez teníamos los datos, había que limpiarlos y organizarlos. Algunos registros estaban incompletos o tenían formatos distintos, así que los normalizamos para que todos los modelos pudieran usarlos sin problemas. Por ejemplo, estandarizamos unidades de distancia, verificamos minutos jugados y rellenamos valores faltantes de manera cuidadosa para que no afectaran las predicciones.

- **Selección y entrenamiento de modelos:**
  Para predecir el valor relativo de los jugadores usamos varios modelos de machine learning:

- **Regresión lineal:**

  Es el modelo más básico y nos sirve para ver cómo las estadísticas principales (goles, asistencias, minutos jugados, duelos ganados, pases clave, etc.) afectan directamente al valor del jugador. Es fácil de interpretar, pero no captura relaciones complejas entre métricas menos evidentes.

- **K-Nearest Neighbors (KNN):**

  Este modelo predice el valor de un jugador comparándolo con los jugadores más parecidos en estadísticas. Es útil para identificar patrones de similitud, aunque puede fallar si el jugador es muy distinto al resto.

- **Árbol de decisión:**

  Divide los jugadores según condiciones en las estadísticas (por ejemplo, número de goles o pases clave) hasta estimar su valor. Es fácil de visualizar y muestra qué métricas son más importantes, aunque un solo árbol puede ser sensible a datos atípicos.

- **Random Forest:**

  Es una versión más avanzada del árbol de decisión que combina muchos árboles diferentes para hacer predicciones más estables y precisas. Es muy útil para detectar patrones complejos y reducir errores que un solo árbol podría cometer.

- Cada modelo se entrenó con datos históricos de jugadores y equipos, y luego se evaluó su precisión comparando las predicciones con datos reales de temporadas anteriores.

- **Evaluación y ajuste de los modelos:**

  Para asegurarnos de que los modelos no solo memorizaban los datos, aplicamos **validación cruzada**, separando los datos en conjuntos de entrenamiento y prueba varias veces. Además, ajustamos los parámetros de cada modelo para mejorar la precisión y evitar errores que puedan dar predicciones poco realistas.

- **Desarrollo de la página web:**

  Con los modelos listos, creamos una web que permite:

  - Consultar estadísticas individuales y comparar jugadores similares.
  - Ver rankings de goleadores, asistentes y defensores destacados.
  - Explorar el rendimiento de los equipos y ver quién rinde mejor o está más cansado.

    Todo se hizo pensando en la **usabilidad y claridad**, usando gráficos interactivos y tablas fáciles de interpretar.

- **Pruebas finales y retroalimentación:**

  Por último, probamos el sistema con datos recientes y pedimos opiniones a compañeros y amigos para mejorar la claridad y la experiencia de uso. Esto nos permitió ajustar la web y los gráficos, haciendo que la herramienta sea intuitiva y útil tanto para entrenadores como para aficionados.

En resumen, la metodología combina **recogida y preparación de datos**, **aprendizaje automático** y **desarrollo web interactivo**, siguiendo un enfoque iterativo donde vamos mejorando los modelos y la interfaz a medida que avanzamos. Esto nos permitió crear un sistema **objetivo, fiable y práctico**, que facilita la evaluación del rendimiento futbolístico de manera clara y directa.

4. Marco normativo

En este proyecto, trabajamos con datos de futbolistas y equipos para analizar su rendimiento y estimar su valor relativo usando modelos de machine learning. Aunque la mayoría de los datos son estadísticos y están disponibles públicamente (como goles, asistencias, minutos jugados, pases clave, duelos ganados, etc.), es importante tener en cuenta la privacidad y protección de datos personales de los jugadores, especialmente si en algún momento se manejan datos identificables o sensibles.

Por eso, este proyecto se ajusta al Reglamento General de Protección de Datos (RGPD) y a la Ley Orgánica 3/2018 de Protección de Datos Personales y Garantía de los Derechos Digitales (LOPDGDD). Ambos marcos normativos establecen que cualquier tratamiento de datos personales debe ser legal, transparente y limitado a lo estrictamente necesario, y que los titulares de los datos tienen derechos que deben respetarse en todo momento.

Tipos de datos tratados

En nuestro caso, los datos que se utilizan son principalmente:

- Estadísticas públicas de rendimiento deportivo: goles, asistencias, minutos jugados, precisión de pases, duelos ganados, distancia recorrida, etc.
- Datos identificativos mínimos: nombre del jugador y equipo al que pertenece
- Estos datos se usan únicamente con fines de análisis estadístico y desarrollo del TFE, y no se emplean para ningún otro propósito comercial o divulgativo fuera del proyecto.

Finalidad del tratamiento

La finalidad del tratamiento de los datos es:

- Realizar análisis de rendimiento individual y colectivo de los jugadores y equipos.
- Estimar un valor relativo de los futbolistas mediante modelos de machine learning. ç
- Permitir la visualización y comparación de jugadores en la herramienta web desarrollada, de manera objetiva y educativa.

En ningún caso se busca difundir datos personales sensibles ni realizar seguimiento fuera del contexto del proyecto académico.

Operaciones de tratamiento realizadas

Las operaciones que se realizan sobre los datos incluyen:

- **Recolección y organización:** recopilación de estadísticas públicas y estructuración en bases de datos para análisis.
- **Procesamiento y análisis:** uso de algoritmos de machine learning (Regresión lineal, KNN, Árbol de decisión y Random Forest) para estimar el valor de los jugadores y generar comparativas.
- **Visualización y presentación:** exportación de resultados en gráficos, tablas y rankings en la página web del proyecto, de manera agregada y sin exponer información sensible.

Medidas para garantizar el cumplimiento

Para cumplir con la normativa y proteger la privacidad de los jugadores, se han tomado varias medidas:

- Se usan solo datos necesarios para el análisis, evitando recolectar información personal sensible que no aporte valor al proyecto.
- Los datos se almacenan en entornos seguros y no se comparten con terceros.
- Se documenta todo el proceso de tratamiento, de manera que se pueda demostrar diligencia y responsabilidad ante cualquier revisión.
- La herramienta web solo muestra datos agregados o estadísticos, garantizando que no se vulneren los derechos de los jugadores.

.

5. Desarrollo específico de la contribución

5.1 Enfoque general del proyecto

La metodología seguida en este proyecto se basa en la combinación de análisis de datos, técnicas de Machine Learning y desarrollo de una aplicación web interactiva. El objetivo principal no ha sido únicamente construir un modelo predictivo, sino desarrollar un sistema completo que permita analizar, interpretar y visualizar el rendimiento de futbolistas de una forma clara y útil.

El enfoque del trabajo es eminentemente práctico. Aunque se han tenido en cuenta metodologías clásicas como CRISP-DM, el desarrollo se ha adaptado a un contexto académico y de aprendizaje, donde el proceso ha sido iterativo. Esto significa que en varias ocasiones fue necesario volver atrás, replantear decisiones tomadas anteriormente y mejorar distintos aspectos del proyecto a medida que se adquiría una mayor comprensión de los datos y de las técnicas utilizadas.

El proyecto se estructura en varias fases claramente diferenciadas:

1. Recopilación y selección del dataset.
2. Limpieza y preparación de los datos.
3. Análisis exploratorio de datos.
4. Creación de nuevas variables.
5. Desarrollo y comparación de modelos de Machine Learning.
6. Visualización de datos mediante Power BI.
7. Desarrollo de una aplicación web interactiva.

Esta estructura ha permitido abordar el problema de manera ordenada y progresiva, asegurando que cada fase se apoye correctamente en la anterior.

5.2 Recopilación de datos

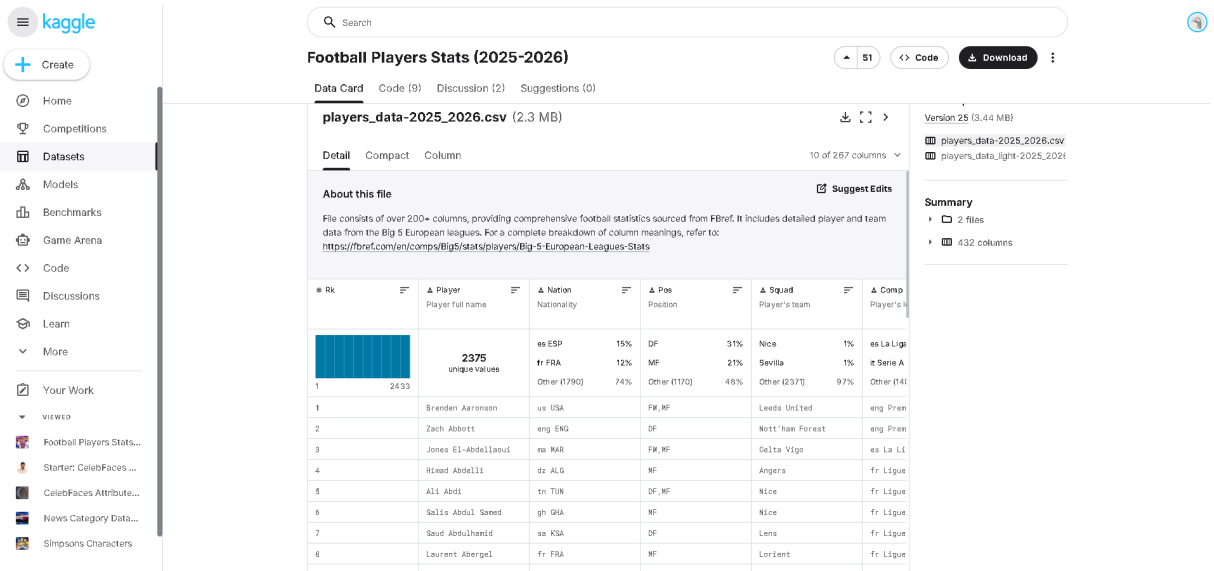
La primera fase del proyecto consistió en la búsqueda de un conjunto de datos adecuado que permitiera trabajar con métricas avanzadas de rendimiento futbolístico. Para ello, se recurrió a la plataforma Kaggle, una de las fuentes más utilizadas en proyectos de análisis de datos y Machine Learning debido a la calidad y variedad de los datasets disponibles.

El dataset seleccionado contiene estadísticas de jugadores de fútbol profesional, incluyendo información tanto a nivel individual como colectivo. Se eligió este conjunto de datos porque incluye métricas avanzadas que van más allá de las estadísticas tradicionales, permitiendo un análisis más profundo del rendimiento real de los futbolistas.

Entre las variables disponibles se encuentran:

- Métricas ofensivas como goles, xG y xAG.
- Variables de creación de juego como pases progresivos y conducciones.
- Indicadores defensivos como entradas e intercepciones.
- Datos contextuales como edad, posición y equipo.

Durante esta fase también se realizó una primera inspección visual del dataset para comprobar su estructura, el número de registros y la coherencia general de las variables. Esta revisión inicial fue fundamental para detectar posibles problemas que posteriormente habría que resolver en la fase de limpieza.



También confiamos en este dataset, ya que la información viene de una fuente fiable que es transfermanket que es una empresa que tiene información de todos los futbolistas del mundo, entonces podíamos fiarnos fácilmente de este dataset

---

5.3 Limpieza de datos y análisis exploratorio

5.3.1 Limpieza de datos

Una vez obtenido el dataset, se procedió a la limpieza de los datos utilizando Jupyter Notebook como entorno principal de trabajo. Esta fase es una de las más importantes del proyecto, ya que la calidad de los modelos y de los análisis posteriores depende directamente de la calidad de los datos de entrada.

El dataset original presentaba varios problemas habituales en datos reales:

- Valores nulos en diferentes columnas.
- Variables con formatos incorrectos.
- Columnas redundantes o irrelevantes.
- Datos que no estaban normalizados.

Las principales acciones realizadas durante la limpieza fueron:

- Eliminación de registros duplicados.
- Sustitución o eliminación de valores nulos, dependiendo de la variable.
- Conversión de columnas a formatos numéricos adecuados.
- Eliminación de columnas que no aportaban valor al análisis.
- Revisión manual de valores extremos que podían distorsionar los resultados.

Este proceso permitió obtener un conjunto de datos limpio, coherente y preparado para su análisis.

1.3) ¿Hay valores nulos o filas duplicadas?

- Calculamos el % de nulos por columna.
- Contamos filas duplicadas exactas.

[5]:
```
# % de valores nulos por columna
na_pct = (df.isna().mean() * 100).sort_values(ascending=False)
print("% de nulos por columna:")
print(na_pct.round(2))

# Filas duplicadas exactas
dup = df.duplicated().sum()
print("\nFilas duplicadas exactas:", dup)
```

```
% de nulos por columna:
CS%          93.98
PKA          93.89
D            93.89
W            93.89
Save%        93.89
             ...
xAG           0.00
npxG+xAG      0.00
PrgC          0.00
PrgP          0.00
Rk            0.00
Length: 267, dtype: float64

Filas duplicadas exactas: 0
```

- Vemos que hay muchos jugadores que tienen valores nulos pero no podemos borrarlo ya que perderiamos informacion porque por ejemplo en la columna save% es para los porteros no para los jugadores entonces por eso hay muhco valores 0 porque no hay nada que poner

```
['Rk_stats_defense', 'Nation_stats_defense', 'Pos_stats_defense', 'Comp_stats_defense', 'Age_stats_defense', 'Born_stats_defense', '90s_stat
s_defense', 'Att_stats_defense', 'Blocks_stats_defense', 'Sh_stats_defense']
```

[33]:
```
informacion_cols = [
    'Player','Nation','Age','Pos','Squad','Comp'
]
participacion_cols = [
    'MP','Starts','Min','90s'
]
progresion_cols = [
    'Touches','Carries','PrgDist','PrgC','PrgP','PrgR'
]
desequilibrio_cols = [
    'Att_stats_possession',   # Regates intentados
    'Succ',                   # Regates completados
    'Succ%'                   # % éxito
]
ofensiva_cols = [
    'Gls','Ast','G+A',
    'xG','npxG','xAG','xG+xAG','Crs'
]
defensa_cols = [
    'Tkl','Int','Tkl+Int','Blocks'
]
eficiencia_cols = [
    'Sh','SoT','SoT%',
    'Sh/90','SoT/90',
    'G/Sh','G/SoT',
    'G-xG','np:G-xG'
]
creacion_cols = [
    'KP','PPA','1/3','CrsPA',
    'SCA','SCA90','GCA','GCA90'
]
contexto_cols = [
    'Att Pen',                  # toques en área rival
    'Def 3rd_stats_possession',
    'Mid 3rd_stats_possession',
    'Att 3rd_stats_possession',
    'Live_stats_possession',
    'Rec',                      # recepciones
    'Won',                      # duelos aéreos ganados
    'Lost',                     # duelos perdidos
    'Cmp',                      # pases completados
    'Crs',                      # centros
    'Clr'                       # despejes
]
```

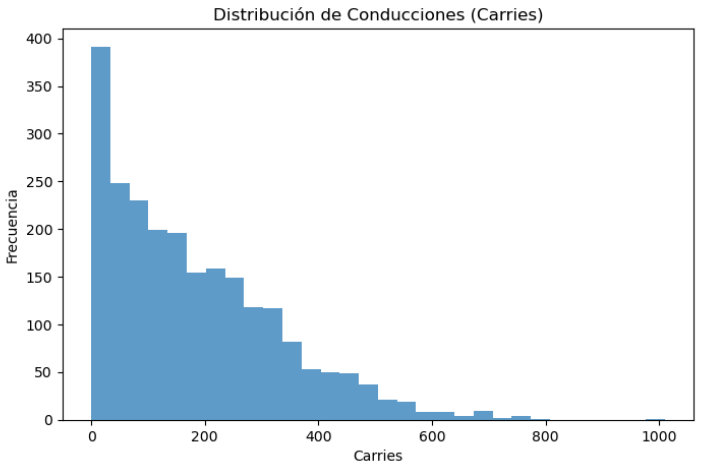
5.3.2 Análisis Exploratorio de Datos (EDA)

Tras la limpieza, se llevó a cabo un análisis exploratorio de datos (EDA) con el objetivo de comprender mejor la información disponible y detectar patrones relevantes. En esta fase se utilizaron gráficos y estadísticas descriptivas para analizar la distribución de las principales métricas.

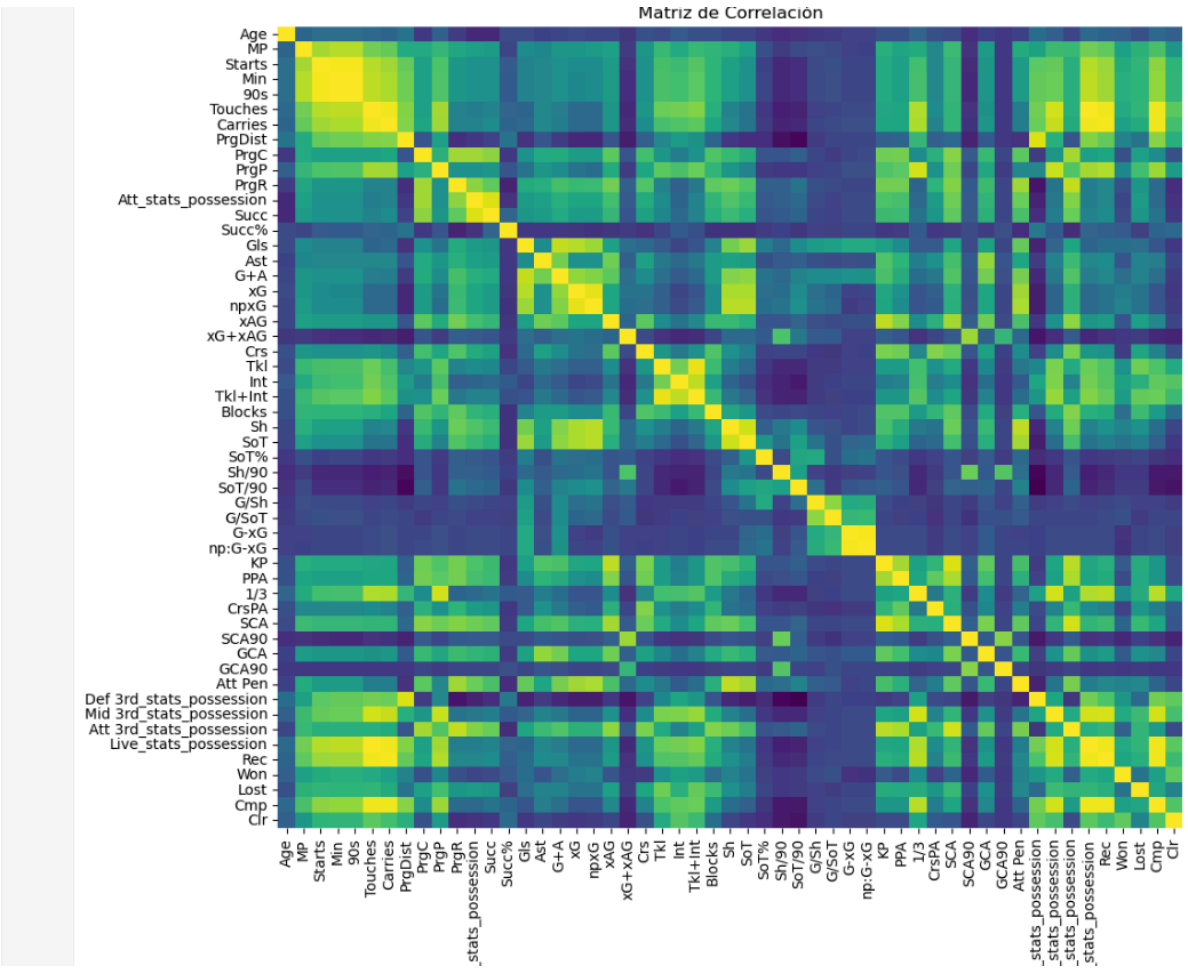
Se estudiaron aspectos como:

- Distribución de métricas ofensivas y defensivas.
- Diferencias de rendimiento entre posiciones.
- Correlaciones entre variables como xG, xAG y goles.
- Relación entre edad y rendimiento.

Este análisis permitió identificar qué variables tenían mayor peso y cuáles podían ser menos relevantes para el modelo predictivo. Además, ayudó a detectar posibles problemas de multicolinealidad entre variables.



- En ste histograma sirve para medir cuántos jugadores progresan conduciendo, vemos que mas del 350 de jugadores no conducen el balon pero es normal ya que hay pocos jugadores que conducen mucho (extremos élite), y esto es clave para el perfilado de jugadores (player profiling).



5.4 Creación de nuevas variables

Una parte clave del proyecto fue la creación de nuevas variables a partir de las métricas originales del dataset. El objetivo de esta fase fue enriquecer la información disponible y capturar mejor el rendimiento real de los jugadores.

Se desarrollaron nuevas columnas como:

- Métricas normalizadas por minutos jugados.
- Índices de rendimiento ofensivo.
- Indicadores de participación en la construcción del juego.

- Un índice de fatiga basado en la carga de minutos y acciones realizadas.

Estas nuevas variables permitieron representar aspectos del rendimiento que no se reflejan directamente en las métricas originales y resultaron especialmente útiles para los modelos de Machine Learning.

5.5 Desarrollo de los modelos de Machine Learning

5.5.1 Definición del problema

El problema abordado se formuló como un problema de regresión supervisada. El objetivo era estimar un valor de mercado relativo de los futbolistas a partir de sus métricas de rendimiento.

Para ello, se dividió el dataset en conjuntos de entrenamiento y test, asegurando una evaluación justa de los modelos.

5.5.2 Modelos implementados

Se implementaron varios modelos con el objetivo de comparar su rendimiento:

- **Regresión Lineal** como modelo base.
- **K-Nearest Neighbors (KNN)** como modelo basado en instancias, sensible a la complejidad del dato y con regularización implícita mediante el parámetro k.
- **Random Forest** como modelo no lineal basado en conjuntos (ensemble) de árboles de decisión.
- **Árbol de Decisión** como modelo no lineal interpretable basado en reglas, que sirve como base para modelos ensemble más avanzados..

Los modelos más complejos demostraron una mayor capacidad para capturar relaciones no lineales entre las variables.

```
Logistic Regression Accuracy: 0.9833679833679834
              precision    recall  f1-score   support

           0       0.97      1.00      0.98       246
           1       1.00      0.97      0.98       235

    accuracy                           0.98       481
   macro avg       0.98      0.98      0.98       481
weighted avg       0.98      0.98      0.98       481

Decision Tree Accuracy: 0.9854469854469855
              precision    recall  f1-score   support

           0       0.99      0.98      0.99       246
           1       0.98      0.99      0.99       235

    accuracy                           0.99       481
   macro avg       0.99      0.99      0.99       481
weighted avg       0.99      0.99      0.99       481

Random Forest Accuracy: 0.9792099792099792
              precision    recall  f1-score   support

           0       0.97      0.99      0.98       246
           1       0.99      0.97      0.98       235

    accuracy                           0.98       481
   macro avg       0.98      0.98      0.98       481
weighted avg       0.98      0.98      0.98       481
```

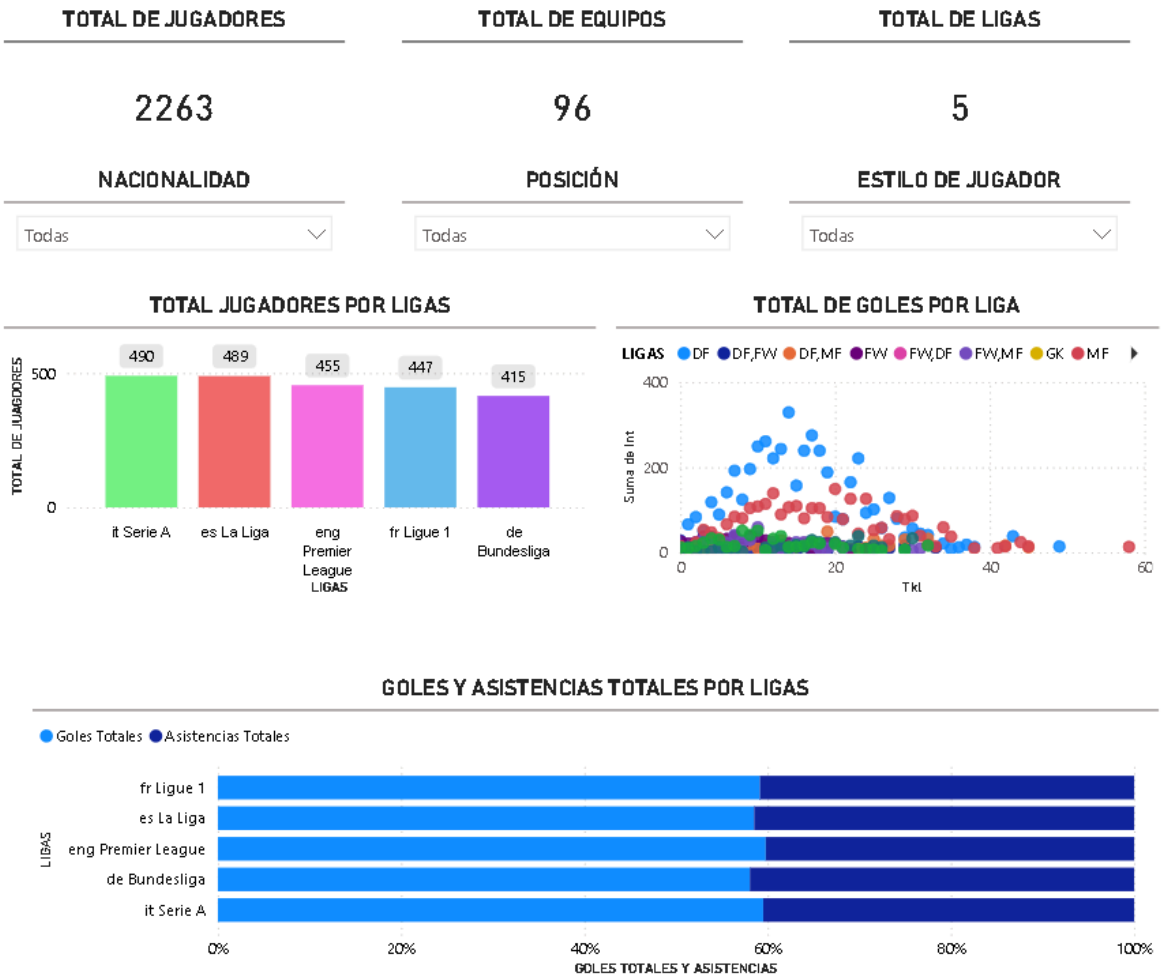
5.6 Visualización con Power BI

Para complementar el análisis realizado en Python, se desarrolló un dashboard interactivo en Power BI. Este dashboard permitió explorar los datos de forma visual y detectar patrones que no siempre son evidentes en análisis numéricos.

Las visualizaciones incluyen:

- Comparativas entre jugadores.
- Análisis por equipos.

- Evolución de métricas clave.



5.7 Desarrollo de la aplicación web

La fase final del proyecto consistió en el desarrollo de una aplicación web interactiva utilizando Streamlit. Esta aplicación permite al usuario explorar los datos, comparar jugadores y equipos y visualizar los resultados del modelo de Machine Learning.

La plataforma integra:

- Visualización de perfiles de jugadores.
- Comparaciones jugador vs jugador.
- Comparaciones equipo vs equipo.
- Gráficos radar interactivos.

Se puso especial atención al diseño visual y a la experiencia de usuario, utilizando animaciones y una estética cuidada.

5.8 Desarrollo de la Inteligencia Artificial

Para la fase de desarrollo de la IA se implementó un sistema que combina modelos de Machine Learning previamente entrenados con una interfaz web interactiva desarrollada en Flet. El objetivo de esta sección no ha sido únicamente predecir métricas de jugadores y resultados de partidos, sino ofrecer un entorno accesible y visual que permita al usuario explorar y comparar información de manera dinámica.

5.8.1 Preparación de datos y creación de métricas adicionales

Se utilizó el dataset previamente limpiado y enriquecido con nuevas variables. Además, se calcularon métricas específicas para cada jugador, como:

- **Potencial**: combinando métricas de creatividad, goles y asistencias ajustadas por edad.
- **Eficiencia ofensiva**: goles por tiros a puerta (*SoT*).
- **Impacto defensivo**: suma de entradas y intercepciones.
- **Score de ataque, defensa y posesión**: agregando métricas de rendimiento clave para simular la contribución al rendimiento del equipo en un partido.

Estas métricas fueron utilizadas tanto para las predicciones individuales como para estimar resultados de enfrentamientos entre equipos.

5.8.2 Integración de modelos de Machine Learning

El sistema incorpora varios modelos entrenados con **Random Forest** para predecir diferentes aspectos del rendimiento futbolístico:

- **Valor de mercado de jugadores de campo y porteros**.
- **Goles y asistencias para jugadores de campo**.
- **Paradas para porteros**.

5.8.3 Funcionalidad de predicción

Se implementaron dos funciones principales:

1. **predecir_jugador(nombre)**: calcula las métricas individuales y predicciones para un jugador específico. Diferencia entre porteros y jugadores de campo, mostrando valor de mercado, goles, asistencias, paradas y fatiga.

2. **predecir_partido(equipoA, equipoB)**: estima probabilidades de victoria basadas en la suma de scores de ataque, defensa y posesión de ambos equipos, generando un análisis interpretativo que resalta fortalezas y debilidades de cada equipo.

5.8.4 Interfaz web con Flet

Se desarrolló una **interfaz interactiva** en Flet que permite:

- Introducir consultas sobre jugadores o partidos.
- Visualizar resultados de manera clara con burbujas de chat diferenciadas por usuario y asistente.
- Analizar estadísticas y predicciones en tiempo real con un diseño visual cuidado, utilizando colores y estilos que mejoran la experiencia del usuario.

6. Código fuente y datos analizados

6.1. Código fuente

Todo el código fuente desarrollado durante la realización de este Trabajo de Fin de Estudios ha sido implementado íntegramente por el autor del proyecto. El desarrollo incluye los procesos de análisis y limpieza de datos, la construcción y evaluación de modelos de Machine Learning, así como el desarrollo de la aplicación web y las visualizaciones asociadas.

Con el objetivo de garantizar la transparencia, la reproducibilidad del trabajo y la correcta evaluación del mismo, el código fuente completo se encuentra alojado en un repositorio público de GitHub, del cual el estudiante es el único autor y propietario. No existen contribuciones externas ni commits realizados por otros usuarios, cumpliendo así los requisitos establecidos para este tipo de trabajos académicos.

El repositorio incluye:

- Notebooks de Jupyter utilizados para la exploración de datos (EDA), limpieza y generación de nuevas variables.
- Scripts de Python correspondientes al entrenamiento y evaluación de los modelos de Machine Learning.

- Código de la aplicación web desarrollada para la visualización e interacción con los resultados.
- Recursos asociados a las visualizaciones y a la presentación de resultados (Power BI y exportaciones).

6.2. Datos Analizados

Los datos utilizados en este proyecto proceden de un conjunto de datos público obtenido a través de la plataforma Kaggle, centrado en estadísticas avanzadas de jugadores de fútbol. Dicho dataset incluye tanto métricas tradicionales como variables avanzadas de rendimiento, lo que lo hace especialmente adecuado para el análisis y la aplicación de técnicas de Machine Learning.

Durante el desarrollo del proyecto, los datos fueron sometidos a un proceso exhaustivo de limpieza, transformación y enriquecimiento, con el fin de garantizar su calidad y coherencia. Este proceso incluyó:

- Eliminación o tratamiento de valores nulos.
- Normalización de variables.
- Creación de nuevas columnas derivadas a partir de métricas existentes.
- Filtrado de registros con información incompleta o poco representativa.

Siempre que ha sido posible, los datos utilizados para el análisis y las versiones procesadas del dataset se han incluido en el mismo repositorio que el código fuente, facilitando la replicabilidad del estudio. En los casos en los que no se ha incluido el dataset completo por razones de tamaño o licencia, se han proporcionado instrucciones claras para su descarga y preparación.

7. Conclusiones

El objetivo principal de este Trabajo de Fin de Estudios ha sido desarrollar un sistema basado en análisis de datos, Machine Learning y visualización interactiva que permita estimar un valor de mercado relativo de futbolistas a partir de métricas avanzadas de rendimiento, así como facilitar la comparación entre jugadores y equipos mediante una plataforma web intuitiva.

El problema abordado parte de una limitación clara en el análisis tradicional del fútbol, donde el rendimiento de los jugadores suele evaluarse mediante estadísticas básicas que no reflejan de forma completa su influencia real en el juego. Además, aunque existen plataformas especializadas, muchas de ellas utilizan modelos opacos y no ofrecen una explicación clara ni personalizable de los resultados.

Para dar respuesta a este problema, el trabajo se ha abordado siguiendo un enfoque práctico y progresivo. En primer lugar, se realizó una recopilación y limpieza exhaustiva de datos procedentes de un dataset público, aplicando técnicas de análisis exploratorio para comprender la estructura y calidad de la información. Posteriormente, se diseñaron y entrenaron distintos modelos de Machine Learning supervisado, comenzando por modelos base como la regresión lineal y avanzando hacia algoritmos más complejos como Random Forest, con el fin de comparar su rendimiento y seleccionar el más adecuado.

Los resultados obtenidos muestran que los modelos no lineales, especialmente Random Forest , ofrecen un mejor ajuste y una mayor capacidad predictiva en la estimación del valor de mercado relativo de los jugadores, confirmando así la hipótesis inicial planteada en los objetivos del trabajo. Estos modelos han demostrado ser más eficaces a la hora de capturar relaciones complejas entre métricas avanzadas de rendimiento.

Además, uno de los aspectos más relevantes del proyecto ha sido la integración de estos modelos en una aplicación web desarrollada específicamente para la visualización y exploración de resultados. Esta plataforma permite comparar jugadores y equipos, visualizar métricas clave mediante gráficos radar y presentar la información de forma clara y accesible, cumpliendo con el objetivo de acercar el análisis avanzado a usuarios no expertos.

En relación con los objetivos planteados inicialmente:

- Se ha conseguido analizar y limpiar un conjunto de datos complejo, generando nuevas variables relevantes.
- Se han implementado y comparado varios modelos de Machine Learning, seleccionando aquellos con mejor rendimiento.
- Se ha desarrollado una plataforma web funcional que integra análisis, predicción y visualización.
- Se ha demostrado la viabilidad de combinar Machine Learning y desarrollo web en un sistema coherente y aplicable al análisis futbolístico.

Por tanto, se puede concluir que los objetivos del trabajo han sido alcanzados de manera satisfactoria, y que la solución propuesta constituye una aportación válida y relevante dentro del ámbito del análisis de datos aplicado al deporte.

8. Limitaciones y prospectiva

8.1. Limitaciones

A pesar de que los resultados obtenidos en el proyecto son bastante buenos y cumplen con los objetivos planteados, es importante reconocer que el trabajo también tiene una serie de limitaciones que conviene mencionar para entender mejor su alcance real.

En primer lugar, los datos utilizados provienen de una única fuente pública. Aunque esta fuente ofrece estadísticas bastante completas y algunas métricas avanzadas, sigue siendo una visión limitada del rendimiento real de los jugadores. No se dispone de información más contextual, como datos tácticos concretos, estado físico detallado, historial de lesiones o incluso factores económicos externos, que en la vida real influyen mucho en el valor de mercado de un futbolista. Al no contar con este tipo de información, el análisis se centra únicamente en el rendimiento estadístico y deja fuera otros aspectos importantes del fútbol profesional.

Otra limitación a tener en cuenta es el tamaño y la distribución del conjunto de datos. No todos los jugadores ni todas las posiciones están igualmente representados, lo que puede

hacer que el modelo funcione mejor para unos perfiles de jugadores que para otros. Además, el valor de mercado que se usa como variable objetivo no es una medida totalmente objetiva, sino una estimación externa basada en distintos criterios. Esto significa que el modelo aprende a predecir un valor que ya tiene cierto margen de error de base, lo que puede afectar a la precisión final de los resultados.

Desde el punto de vista del modelado, aunque se han probado varios algoritmos de machine learning como regresión lineal, KNN, árbol de decisión y Random Forest, no se ha realizado una optimización muy profunda de los hiperparámetros. Tampoco se han explorado modelos más avanzados, como redes neuronales profundas, principalmente por limitaciones de tiempo y de recursos computacionales. Aun así, los modelos utilizados han permitido obtener resultados coherentes y comparar diferentes enfoques de predicción.

Por último, la aplicación web desarrollada cumple bien su función de análisis y visualización, pero tiene un enfoque principalmente exploratorio y muy limitada ya que actualmente no cuenta con características más avanzadas no tiene filtros por decirlos asi es una pagina bonita pero no es como muy profesional

8.2. Trabajo futuro

El trabajo desarrollado abre múltiples líneas de mejora y ampliación que podrían aportar un valor añadido significativo al sistema propuesto.

Una primera línea de trabajo futuro sería la ampliación del conjunto de datos, incorporando información procedente de múltiples temporadas, diferentes ligas y fuentes adicionales. Esto permitiría entrenar modelos más robustos y generalizables, así como analizar la evolución temporal del rendimiento de los jugadores.

En el ámbito del Machine Learning, se podrían explorar modelos más avanzados, como redes neuronales, con el objetivo de mejorar la precisión de las predicciones y aumentar la interpretabilidad de los resultados. Asimismo, una optimización más profunda de hiperparámetros podría contribuir a mejorar el rendimiento de los modelos actuales.

Otra línea interesante sería la integración de datos en tiempo real y la automatización del pipeline de datos, permitiendo actualizar las predicciones de forma continua a medida que se generan nuevos datos de partidos.

Desde el punto de vista del desarrollo web, el sistema podría evolucionar hacia una plataforma más completa, incorporando perfiles de usuario, comparativas personalizadas,

exportación de informes y despliegue en la nube. Esto facilitaría su uso en contextos profesionales, como scouting, análisis técnico o toma de decisiones deportivas.

También podríamos escalarlo a todas las ligas del mundo, recordamos que solo lo hemos hecho de las 5 grandes ligas, pero podríamos hacerlo a muchas más para que sea para todos los entrenadores y staff, es algo necesario y que puede ayudarlos

Referencias bibliográficas

Aquí tenemos la bibliografía que me ha ayudado a buscar información sobre mi proyecto

para saber si podía ser importante para las personas o que pueda a llegar a funcionar.

Álvarez, J., & Blanco, A. (2020). Aplicación del análisis de datos al rendimiento deportivo en el fútbol profesional. *Revista Internacional de Ciencias del Deporte (RICYDE)*, 16(62), 420–436

Barajas, Á., & Rodríguez, P. (2014). Economía del fútbol: una visión global. *Revista de Economía Aplicada*, 22(65), 5–32.

Benítez, J. M., Castro, J. L., & Requena, I. (2010). Técnicas de minería de datos aplicadas al deporte. *Revista Iberoamericana de Inteligencia Artificial*, 14(46), 35–48.

Carling, C., Williams, A. M., & Reilly, T. (2005). *Análisis del rendimiento en el fútbol*. Barcelona: Editorial Paidotribo.

García, J., Ibáñez, S. J., & Feu, S. (2011). Análisis del rendimiento en deportes colectivos: una revisión metodológica. *Cuadernos de Psicología del Deporte*, 11(2), 45–60.

Gómez, M. A., Lago-Peñas, C., & Pollard, R. (2013). Situational variables and physical performance in soccer. *Journal of Human Kinetics*, 35, 123–133.
(Artículo disponible en español en versiones académicas)

Hernández Orallo, J., Ramírez Quintana, M. J., & Ferri Ramírez, C. (2011). *Introducción a la minería de datos*. Madrid: Pearson Educación.

López Peña, J. L., & Sánchez, F. (2019). Big Data y analítica avanzada aplicada al fútbol profesional. *Cuadernos de Entrenamiento Deportivo*, 4(1), 15–29.

Mallo, J., & Navarro, E. (2008). Análisis del esfuerzo físico en el fútbol mediante tecnología GPS. *Apunts. Educación Física y Deportes*, 92, 14–22.

Méndez, C., & García, F. (2021). Modelos predictivos aplicados al deporte mediante aprendizaje automático. *Revista Iberoamericana de Sistemas, Cibernética e Informática*, 18(1), 52–61.

Pérez, J., & Castellano, J. (2015). Uso de indicadores de rendimiento para la evaluación del juego en fútbol. *Revista de Psicología del Deporte*, 24(1), 121–128.

Perarnau, M. (2016). *La evolución táctica del fútbol moderno*. Barcelona: Editorial Roca.

Sánchez-Sánchez, J., & Burillo, P. (2016). Análisis de métricas físicas y técnicas en fútbol de alto nivel. *Retos: Nuevas Tendencias en Educación Física, Deporte y Recreación*, 30, 252–257.

Torres, D., & López, J. (2020). Visualización de datos aplicada al análisis deportivo. *Revista Española de Documentación Científica*, 43(3), e270.