TFM – Máster en Data Analytics

VCF Forecasting Scouting Model

Aplicación para la predicción y visualización del valor de mercado y el rendimiento de los jugadores.

Autor: Cristian Marty, Andrés Cervera, Jesús Jornet, Juan Cornejo.

Tutor: Pedro Nieto

Septiembre 2024

AGRADECIMIENTOS:

Este Trabajo de Fin de Máster ha sido un desafío tanto personal como académico, y no habría sido posible sin el apoyo de muchas personas a las que nos gustaría expresar nuestro más sincero agradecimiento.

Desde nuestros tutores y mentores Pedro, Victoria, Javier, Miguel, Rafa y Jose, hasta todos y cada uno de los docentes que han pasado por nuestra aula en cada uno de los módulos. Gracias por cada una de las piedras que han puesto en nuestra formación hemos sido capaces de desarrollar con éxito este proyecto, aplicando todo lo aprendido en este maravilloso año lleno de experiencias.

TABLA DE CONTENIDO

1. ABSTRACT Y RESUMEN

1.1. VERSIÓN EN CASTELLANO

En un contexto de máxima exigencia como el fútbol de élite, el cuál actualmente se está viendo afectado por algunas constricciones y limitaciones financieras, cualquier pequeño detalle puede contribuir a marcar la diferencia.

Es por ello, que este trabajo se centra principalmente en el diseño de una solución que permita al Valencia Club de Fútbol disponer de conocimientos de mercado y, en particular, poder disponer de un método de estimación de precios de transferencia de los derechos federativos y económicos en las operaciones de traspaso mediante algoritmos de machine learning, con el objetivo de poder alcanzar una ventaja competitiva frente al resto de equipos.

A tal efecto, la solución propuesta está basada en todas las metodologías, herramientas y recursos adquiridos en el Máster de Data Analytics con el fin de poder demostrar el buen aprendizaje y el conocimiento de qué herramienta es más adecuada para cada caso de uso.

Adicionalmente, el equipo ha prestado especial énfasis a la simplificación de la experiencia de usuario de la herramienta. Creemos que una buena interfaz y un proceso fácil y limpio pueden ayudar al uso de la misma, y más en un sector que habitualmente suele estar algo más alejado del ámbito tecnológico. Esta búsqueda se puede ver plasmada en pequeños detalles que serán analizados y desglosados más adelante.

Palabras clave: Transferencia, Fútbol, Machine Learning, Experiencia de Usuario (UX)

1.2. VERSIÓN EN INGLÉS

In a highly demanding context such as elite soccer, which is currently being affected by some financial constraints and limitations, any small detail can make a difference.

That is why this work is mainly focused on the design of a solution that allows Valencia Club de Fútbol to have market knowledge and, in particular, to have a method for estimating transfer prices of federative and economic rights in transfer operations using machine learning algorithms, in order to achieve a competitive advantage over other teams.

To this end, the proposed solution is based on all the methodologies, tools and resources acquired in the Master of Data Analytics in order to demonstrate good learning and knowledge of which tool is best suited for each use case.

Additionally, the team has placed special emphasis on simplifying the user experience of the tool. We believe that a good interface and an easy and clean process can help the use of the tool, especially in a sector that usually tends to be somewhat more distant from the technological field. This search can be seen in small details that will be analyzed and broken down later.

Keywords: Transfer, Football, Machine Learning, User Experience (UX)

2. INTRODUCCIÓN

El objetivo de este TFM es poder ofrecer al Valencia C.F. una solución que le permita estimar precios de traspaso y poder obtener información que le pueda ayudar a tener una ventaja competitiva en el mundo del fútbol.

Sin embargo, para ello, antes es necesario conocer un poco más la idiosincrasia del propio sector, a fin de entender que rol jugará este TFM en la cobertura de las necesidades del club.

2.1. INTRODUCCIÓN AL SECTOR

El fútbol es uno de los deportes más populares en todo el mundo. Este deporte, tan seguido y practicado por tanta gente, presenta una de las paradojas más importantes del deporte. A pesar de que sus reglas básicas son muy sencillas, una vez comienza el partido, el desarrollo del juego es muy complejo.

Esta complejidad del fútbol hace que muchos autores hayan intentado desgranar o clasificar todos los aspectos que tienen relevancia en un partido (táctico, técnico, físico, etc.) a fin de entender cada uno de ellos no sólo de forma individual sino también la correlación entre ellos. De hecho, esta división en tres partes se puede apreciar en la mayoría de literatura de este sector, y en incluso en los cursos de formación de entrenadores, la mayor parte del tiempo se dedica únicamente a estas tres ramas.

Sin embargo, uno de los grandes cambios que ha experimentado el fútbol en los últimos años ha sido la **evolución de la tecnología** y su **implantación** a lo largo de las estructuras de los clubes. Esta evolución se desglosa en el siguiente epígrafe.

2.2. EVOLUCIÓN TECNOLÓGICA EN EL SECTOR

Vivimos en un entorno en el cual muchas empresas y organizaciones, gracias a la ayuda de la tecnología, tienen la posibilidad de recabar una inmensa cantidad de datos y procesarlos, con el fin de extraer conclusiones que les permitan tomar las mejores decisiones posibles para alcanzar sus objetivos.

Dentro de este entorno encontramos el mundo del deporte un sector que genera miles de millones en todo el mundo y en el cual los propios deportistas buscan alcanzar el éxito mediante disciplina, sacrificio y entrenamiento.

El fútbol, en particular, siempre ha sido uno de los deportes que más ha tardado en abrirse a nuevas tecnologías y ha ido evolucionando tanto sus reglas como la forma de practicarlo siempre de forma paulatina.

A pesar de ello, la esencia del fútbol a priori sigue siendo bastante básica: se enfrentan dos equipos de once jugadores cada uno y el objetivo es marcar más goles que el rival. No obstante, una vez comienza el partido, este juego de reglas sencillas se transforma en un juego de desarrollo altamente complejo en el cual pueden influir múltiples variables. Esta característica recuerda un poco al experimento de John Conway y su juego "Game of Life" publicado en el artículo "*Mathematical Games*" (Gardner M. , 1970) en el cual se reflejaba cómo un juego con reglas sencillas puede tener desarrollos muy complejos.

Esta percepción de aparente simpleza podría ser el origen de que mucha gente piense que es fácil entender cómo funciona este deporte (recordándonos al famoso efecto Dunning-Kruger (Dunning, 2011)) y, por tanto, no ven la necesidad de aplicar de estas nuevas tendencias y tecnologías que incluso ya se han empezado a probar en otros deportes de gran prestigio como el baloncesto, el baseball o el fútbol americano.

Hay múltiples ejemplos de cómo estos deportes (con gran repercusión en EEUU) han mostrado una apertura mayor a nuevas tendencias. Uno de ellos podemos encontrarlo en el baseball, y más concretamente en los Oakland Athletics. Gracias al uso de datos de los jugadores, fueron capaces de confeccionar una plantilla de bajo coste que a la postre acabó dando un gran rendimiento. Dicha hazaña, que fue recogida en el famoso libro de

Moneyball (Lewis, 2004), tuvo tanta repercusión que la NBA y la NFL no tardaron en empezar a utilizar datos de los jugadores para confeccionar sus plantillas.

El uso de herramientas de recogida y análisis de grandes volúmenes de datos es utilizado por la gran mayoría de las empresas más potentes del mercado con múltiples objetivos: reducir costes, detectar anomalías en los procesos, identificar patrones de consumo de los clientes etc. De la misma forma, un club de fútbol podría utilizar dicha tecnología no sólo para su área directiva o de gestión, sino también para el ámbito deportivo, pudiendo ser una herramienta muy efectiva no sólo para la mejora de entrenamientos y toma de decisiones, **sino también para la identificación de jugadores con potencial que todavía no hayan sido descubiertos**.

Importancia especial cobra esta última posibilidad. Durante los últimos años, el mercado de traspasos de futbolistas ha experimentado una inflación sin precedentes tal y como muestra la Figura 1, debido principalmente al aumento del valor de venta de los derechos televisivos de los clubs y en algunos casos a la entrada de capital extranjero y fondos de inversión en los clubs.

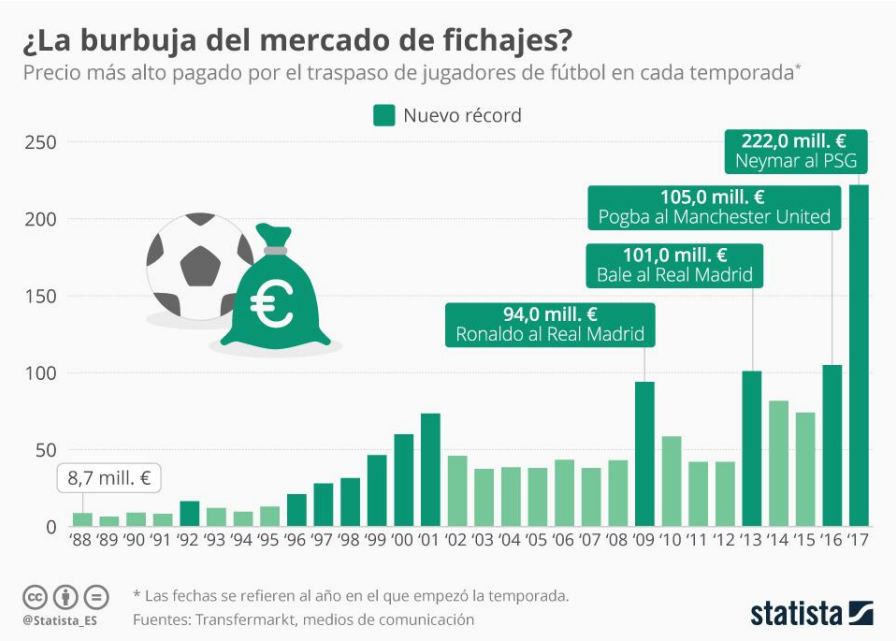

Figura 1. ***Evolución del importe del traspaso más caro de cada año 1988-2017***

Gráfico que muestra la evolución del máximo importe pagado por el traspaso de un jugador durante esa temporada (Moreno, 2017). *Recuperado de la web: https://es.statista.com/grafico/10537/neymar-lleva-al-precio-por-traspaso-a-un-record-historico/*

En consecuencia, el hecho de utilizar inteligentemente la información podría ayudar a los equipos a ser más precisos en la sección de un futbolista y en la predicción de su potencial y su valor. De esta forma, podrían anticiparse al mercado y por lo tanto obtener un mejor precio por el futbolista, con dos principales ventajas como consecuencia:

1) Mayor precisión en la captación de talento
2) Mejor rendimiento del capital disponible para traspasos.

Por estos motivos, el uso de la tecnología de recogida y proceso de datos puede jugar un papel fundamental no sólo en la toma de decisiones en el partido o la temporada, sino también durante el periodo de traspasos.

Sin embargo, en el fútbol siempre ha habido cierto escepticismo en el uso de los datos como una herramienta más. A pesar de que en los últimos años ese enfoque ha cambiado y cada vez más clubs apuestan por sistemas y especialistas de datos, todavía hay muchos profesionales que se muestran reluctantes a esta nueva herramienta.

Algunos de estos casos en los que se muestra la desconfianza en los datos podemos encontrarlos en el libro "*The numbers game: why everything you know about soccer is wrong"* (Anderson & Sally, 2013).

Hasta hace tan sólo una década atrás la recogida de información estadística y visual se realizaba de forma manual. Esto hacía que, en muchos casos, el volumen de datos a recabar fuese limitado y susceptible de errores manuales (errores de conteo, criterios heterogéneos según cada observador, etc.). En consecuencia, la relevancia que se le daba al análisis integral de la información era generalmente bajo.

A pesar de ello, en los últimos años, múltiples clubes de la elite profesional han decidido apostar por sistemas que permitan la fácil captura de información de partidos y jugadores, así como por profesionales que sean capaces de interpretar los resultados obtenidos. Gracias a ello, el cuerpo técnico puede disponer de información en tiempo real acerca de la actuación de los jugadores en el partido y la temporada, además de conocer detalles de jugadores que puedan ser considerados fichajes potenciales que permitan al club anticiparse al mercado, pudiendo suponer un ahorro en costes.

2.3. HISTORIA Y CONTEXTO DEL CLUB

Valencia Club de Fútbol es un club de fútbol histórico de la Liga Española. Este club, que ya es centenario (creado un 1 de marzo de 1919), se ha ganado a pulso ser uno de los referentes en el panorama futbolístico tanto nacional como internacional.

En el ámbito nacional, desde su fundación, salvo en dos breves etapas en las que jugó en segunda división (1928-31 y 1986-87), el equipo siempre se ha mantenido en la división de oro del fútbol español. De hecho, hasta en 6 ocasiones ha levantado el título de campeón de liga y en 8 ocasiones conquistó el título de campeón de Copa, siendo la más reciente esta pasada temporada 2018-19 tal y como se puede ver en la propia web del equipo. (Valencia C.F., 2019)

En el ámbito internacional también ha demostrado su poderío y casta alzando tres títulos europeos como son las Copas de la UEFA cosechadas en las temporadas 1961-62,62-63 y 2003-04. Adicionalmente, también fue dos veces finalista de la UEFA Champions League de forma consecutiva en las temporadas 1999-00 y 2000-01, quedando patente la importancia de este equipo en el panorama futbolístico internacional.

Sin embargo, en las últimas temporadas el club se ha visto afectado por la crisis financiera del COVID. Esto justo con un claro decrecimiento general en los beneficios de los clubes deportivos de la primera división española, ha afectado a la capacidad de maniobra del propio club.

3. OBJETIVOS Y CONTEXTO

Una vez entendido el contexto del sector y la historia del propio club, procederemos a detallar los objetivos de este trabajo de fin de máster (en adelante, “tfm”).

3.1. OBJETIVO GENERAL DEL TFM

Poder proporcionar una solución al Valencia C.F. que le permita hacer una estimación de los precios de transferencia de los jugadores de forma simple y sencilla, a fin de poder obtener una ventaja competitiva en su sector.

3.2. CONTEXTO DEL DESAFÍO Y MOTIVACIÓN

Como se ha analizado en las secciones previas, el mundo del fútbol se ha vuelto cada vez más competitivo. La llegada (e implantación paulatina) de nuevas de tecnologías ha permitido a muchos clubes comenzar a desarrollar nuevas metodologías de trabajo, así como poder invertir más en I+D tanto en los procesos de captación de jugadores, como en los de análisis de rendimiento.

De la misma forma, la viabilidad y sostenibilidad económica se ha vuelto un aspecto fundamental de los clubes de fútbol. Para entender la relevancia de este tfm, es importante entender que, actualmente, aunque los derechos de retransmisión son la principal fuente de ingresos de los clubes de fútbol profesional, los beneficios por traspasos juegan un papel sumamente importante en el balance económico de un club.

Es por eso que, contar con buenas estimaciones que permitan establecer un correcto precio de transferencia, es clave tanto para las compras, como para las ventas. Un indicador como este puede tener un impacto directo tanto en el aumento de ingresos (uso de la estimación para ajustar el precio de venta), como para la reducción de costes (uso de la estimación para ajustar el precio de compra). Por tanto, este tfm puede potencialmente ayudar al club en su gestión y tener un impacto directo en su economía.

3.3. DESAFÍOS PRINCIPALES

Como se ha indicado en el apartado de objetivo general, se busca proveer de una estimación. Es por ello que, aunque el resto de elementos del proyecto puedan jugar un papel relevante, el protagonismo se encuentra en la propia estimación.

Parece lógico pensar entonces que el gran desafío se encontrará en lo relativo a la creación y optimización del proceso de estimación.

A pesar de contar con un proveedor de datos de calidad, es aquí donde podemos encontrar uno de los mayores desafíos: la calidad y cantidad de la información para generar la estimación.

Algunas de las limitaciones identificadas:

- **Información sesgada**: Información ya refinada, es decir, no viene en crudo ("raw") sino que puede haber sido previamente preprocesada por el proveedor de datos, lo que puede otorgar algo de confusión al no saber cómo se ha construido esa ratio. Ejemplo: ELO de un equipo.
- **Alcance de datos limitado**: No se dispone de algunos datos más complejos, o de datos relativos a factores psicológicos o de toma de decisiones que pueden influir en el rendimiento de un jugador y, por tanto, en su posible valor de transferencia.
- **Acceso a la información:** El proveedor de datos no dispone de una API para conexión directa con los datos necesitados. La descarga de datos se realiza de forma manual y en caso de necesitar nuevas temporadas, se realiza una petición al proveedor.
- **Variables con gran cantidad de valores –** Este ha sido un desafío interesante, especialmente a la hora de enfrentar el modelo de ML. Hay algunas variables, como "Nacionalidad" que creemos podrían jugar un papel relevante en la determinación del precio de transferencia. Saber cómo poder procesar una variable de carácter cualitativo con tantas posibles opciones (hasta 195 nacionalidades distintas), sin duda fue uno de los retos a los que nos enfrentamos.

- **Volumen de datos limitado**: No disponemos de históricos de muchos años, tan sólo de algunas temporadas.
- **Distribución de los precios de transferencia:** Parece lógico intuir que la mayoría de traspasos en el mundo del fútbol se realizan por importes pequeños. Esto muestra que no siguen una distribución Normal, y por tanto el número de casos según el importe del traspaso puede mostrar desequilibrios (habiendo muchos traspasos bajos, y pocos traspasos altos). Esto puede tener un impacto a la hora de entrenar el algoritmo, ya que tiene menos casos de los que aprender y, por tanto, puede presentar más dificultades para predecir de forma correcta.

Debido a estas limitaciones, la creación de un proceso de estimación que ofrezca unas garantías mínimas se presenta como uno de los principales desafíos del tfm.

4. METODOLOGÍA

En este apartado procederemos a definir la metodología de trabajo definida para alcanzar el objetivo principal.

4.1. ASIGNACIONES

El grupo está compuesto por cuatro integrantes: Cristian Marty, Juan Cornejo, Andrés Cervera y Jesús Jornet. De cara a afrontar el reto que suponía este tfm; se estableció el siguiente plan de acción:

1) Primera reunión para definir los objetivos y alinearse en cuanto a piezas y estructura propuesta.
2) Asignación principal de áreas:
    a. Cristian Marty: Por su gran conocimiento técnico y experiencia, se encargó principalmente de crear el caparazón que sirve como base para los desarrollos de la estructura (esto es, creación de la base en GCP y streamlit). De la misma forma, participó en la creación de los Cloud Run. Adicionalmente, haría de soporte técnico para el resto de áreas.
    b. Andrés Cervera – Centrado principalmente en el diseño y creación de la API de conexión del modelo, así como en dar soporte en el área de streamlit.
    c. Jesús Jornet – Referente en el área de UX, se centró principalmente en que la interfaz para el usuario fuese sencilla y tuviese sentido/criterio.
    d. Juan Cornejo – Por su experiencia en el sector del fútbol, se encargó del modelado de ML y del proceso para aplicarlo a los datos disponibles en BigQuery.
3) Reuniones grupales periódicas para evaluar la situación del TFM
4) Reuniones parciales entre miembros con piezas que dependen unas de otras para coordinar la evolución de las mismas.

4.2. MUESTRA UTILIZADA

La muestra de datos que se ha utilizado en el tfm está compuesta por un conjunto de jugadores de diversas ligas y temporadas, incluyendo estadísticas clave como la cantidad de goles, asistencias, partidos jugados, edad, posición, y sus valores de mercado en diferentes momentos. El listado de jugadores y sus datos se cargan a través de una aplicación de Streamlit, lo que facilita la interacción con los usuarios. Los datos son procesados y almacenados en BigQuery para su análisis y procesamiento.

Para ser más precisos; los registros de traspasos que componen la muestra constan de los siguientes criterios:

- Traspaso realizado en las últimas 6 temporadas (excluyendo los de este mercado de verano 24/25).
- La Liga de destino fuese una de las 5 grandes ligas
- Traspasos de 1m€ o superior

Tras ese filtrado, la muestra final con la que entrenó el modelo consta del orden de tan solo unos 2300 registros aproximadamente (2.289 para ser exactos).

4.3. PROCEDIMIENTO

Se ha diseñado un sistema automatizado en el que los usuarios cargan un archivo **Excel** con las estadísticas de los jugadores a través de una interfaz web construida con **Streamlit**. Estos datos son almacenados en **BigQuery** y se utiliza un modelo previamente entrenado para hacer las predicciones de los precios de traspaso. El modelo emplea técnicas de machine learning para evaluar múltiples variables, como el rendimiento de los jugadores, su edad, su nacionalidad, entre otros.

A partir de las estadísticas individuales de cada jugador, el sistema predice cuál debería ser su precio de traspaso en el mercado actual. La predicción es procesada a través de un modelo **PKL** alojado en un bucket de GCP, utilizando una API basada en **Flask**. El resultado final se muestra nuevamente en **Streamlit**, donde se presenta al usuario la predicción del precio de traspaso del jugador.

5. DESARROLLO

El desarrollo del presente proyecto se ha estructurado en varias fases claramente definidas, cada una con objetivos y tareas específicas. A continuación, se presenta la composición final del desarrollo del proyecto, dividido en las distintas partes que lo conforman:

5.1. DESPLIEGUE Y VISUALIZACIÓN EN STREAMLIT

La fase de visualización en **Streamlit** permite que el usuario cargue nuevos archivos de datos en formato Excel y vea, tras el procesamiento, el precio de traspaso estimado para cada jugador. Los resultados de las predicciones son mostrados de forma clara e interactiva, permitiendo al usuario evaluar las estimaciones en función de las estadísticas de cada jugador.

Adicionalmente, se incluyó un script para que, al cargar la información del excel se lance un proceso de scrapping para obtener la imagen del jugador (en caso de que la hubiere) por medio de BeatifulSoap. Esto da una capa más de cercanía y experiencia de usuario a la herramienta.

Además, la interfaz incluye funcionalidades para filtrar y ordenar los jugadores por diversos parámetros como el equipo, la posición o la nacionalidad.

5.2. CARGA DE DATOS EN BIGQUERY

Tras la carga de datos a través de streamlit, el archivo de datos es cargado en un bucket de **Google Cloud Storage**. Desde allí, el canal de mensajería de Pub/sub es usado para activar el **Cloud Run**. Esto hace que se ejecute un proceso automatizado para almacenar la información (de forma aditiva, no remplaza lo ya existente) en **BigQuery**, la base de datos de almacenamiento en la nube. Este proceso aseguró la persistencia de los datos y su disponibilidad para las siguientes fases del proyecto.

5.3. ENTRENAMIENTO Y APLICACIÓN DEL MODELO DE PREDICCIÓN

Este desarrollo comprende varias subfases:

ANÁLISIS DE INFORMACIÓN DISPONIBLE

El primer punto de partida. Se realizó un análisis exhaustivo de qué variables se disponían, y si tenían sentido tanto desde un nivel técnico como desde un nivel deportivo. Tras este análisis, se decidió contactar con el proveedor para solicitar más información como la nacionalidad.

ANÁLISIS DE VARIABLES

En este apartado se revisó la información ya definida. Esto implica la búsqueda de valores anómalos o vacíos que pudiesen impactar en el desarrollo de los modelos. De la misma forma, se analizó la distribución de cada una de las variables para identificar posibles complicaciones. No obstante, dado que los datos venían de un proveedor oficial del club, la mayoría de los datos venían limpiados y listos para ser pre-procesados.

Tras el primer análisis, vimos que el campo “Coste” que actuaría como Y a estimar en el modelo, tenía una distribución más similar a una distribución exponencial, donde los valores bajos son mucho más comunes que los altos. Esto es algo que tuvimos que tener presente a la hora de generar el modelo.

PRE-PROCESADO

Encontramos otro desafío importante en las variables cualitativas. Algunas de ellas presentaban un número de valores sumamente alto, como el caso de nacionalidad. Debido a esto, pensamos que la técnica de one-hot encoding no lograría una buena eficiencia ya que aumentaría en gran medida el número de columnas a procesar. Es por ello que optamos por la técnica de embedding para las variables cualitativas, convirtiendo sus valores en vectores de longitud 5.

Posteriormente, se normalizaron los datos utilizando el criterio de MinMax.

GENERACIÓN DEL MODELO

A la hora de crear el modelo, se han probado tanto diferentes enfoques como modelos en sí.

Al principio, se optó por probar cómo de preciso sería un modelo de ML de regresión para estimar la cantidad exacta del traspaso de un jugador. A pesar de probar diferentes modelos de regresión (xGBoost, RF), e incluso redes neuronales, la precisión no fue ideal, siendo en muchos casos baja y teniendo el modelo dificultades para aprender.

Esto lo achacamos principalmente a:

- La distribución del precio de traspasos: Habiendo un alto volumen de traspasos de bajo valor, y pocos casos de alto. Al haber pocos casos de altos valores, al modelo le costaba especialmente predecir bien esos casos.
- Limitación de datos: Seguramente hay más factores que puedan jugar un papel clave en el precio de traspaso de un jugador, como son datos de eventing (pases, centros, remates) como datos físicos (velocidad máxima, esfuerzos a alta intensidad). Al no disponer de esta información, es probable que, aunque se presentase un modelo relativamente sólido y bien estructurado, este no fuese capaz de captar todas las relaciones entre variables.

Debido a esto, optamos por la estrategia de cambiar el enfoque y plantear un algoritmo de ML de clasificación. Para ello, crearíamos diversos tramos de precios de transferencia.

Un ejemplo:

TRAMO	IMPORTE ESTIMADO
1	De 0 a 15m€
2	De 15m€ a 30m€
3	De 30m€ a 60m€
4	+60m€

Al definir tramos relativamente amplios (diferencias de 15m€), ampliamos más el margen de error del algoritmo y le facilitamos la primera identificación del caso.

Una vez hecho esto, probamos diferentes modelos de clasificación y ajuste de hiperparámetros, siendo el Random Forest Classifier el que mejor funcionó, alcanzando

un 82% de precisión en el test (y un 95% en pruebas cruzadas). Una comprobado y testado el modelo, se exportó en un PKL.

Sí cabe destacar que, ese porcentaje se alcanzó al haber aplicado la técnica SMOTE para hacer un oversampling de las clases minoritarias. El objetivo principal era que hubiesen casos suficientes como para que el algoritmo fuese capaz de identificar bien cada uno de los casos.

DESARROLLO DE CLOUD RUN PARA APLICACIÓN DEL MODELO

En esta fase, se desarrolló un servicio en **Cloud Run** para el entrenamiento del modelo de predicción de precios de traspaso de jugadores.

La función de Cloud Run realiza lo siguiente:

Lo primero que hace es coger los datos actuales de BigQuery y los preprocesa en un dataframe (al vuelo) para que tenga la misma estructura que la requerida por el modelo (embedding, droppeo de variables, normalización, etc.).

Posteriormente, busca en un Bucket de GCP el último modelo cargado que haya y le aplica la predicción.

Finalmente, se escribe en la columna “Predicción” de Bigquery.

5.4. DESARROLLO DE LA API DE COMUNICACIÓN

Para intermediar entre la interfaz de **Streamlit** y el modelo entrenado, se implementó una **API REST** utilizando **Flask**. La API permite a Streamlit enviar los datos preprocesados, recibir las predicciones generadas por el modelo y devolver los resultados para ser mostrados al usuario de manera interactiva en la interfaz de la aplicación.

Esta API sirve como puente entre el modelo de machine learning y la visualización en Streamlit, gestionando de manera eficiente las peticiones y respuestas en tiempo real.

5.5. FLUJOGRAMA DEL PROCESO

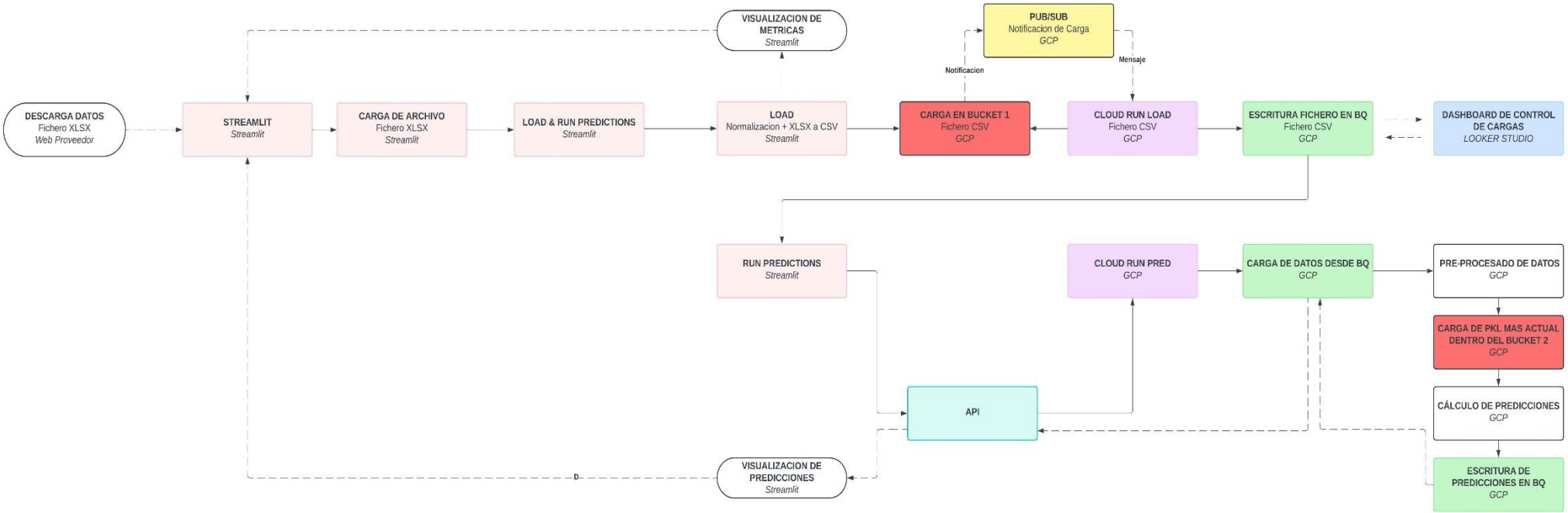

Figura 2. ***Flujograma de la solución diseñada.***

Gráfico que muestran las partes del proceso. Creación propia.

6. DISCUSIÓN

Tras haber analizado la metodología de trabajo y habiendo completado un MVP (Mínimo producto viable) de la propia solución, conviene analizar los resultados a fin de vislumbrar si se han alcanzado los objetivos propuestos.

Para ello, se analizará de forma globa y por partes.

6.1. CARGA DE DATOS

Actualmente, la carga de datos se realiza de forma manual. El usuario ha de descargar de su proveedor de datos un fichero con los datos, en formato xslsx. Esta ha sido una de las limitaciones y puntos a mejorar de cara al futuro.

La carga manual plantea grandes dificultades al usuario. Algunas de ellas son:

- **Dependiente de la estructura del fichero de carga** – En caso de haber alguna modificación en la estructura, puede impactar en todo el proceso e inutilizarlo por completo.
- **Lentitud en la subida de información, no se dispone de información actualizada en tiempo real** – Aunque a priori podría no tener excesivo impacto (ya que no hay requerimiento de instantaneidad en la información), es cierto que, en caso de querer subir grandes volúmenes de información, puede requerir de tiempo.
- **Dificultad para modificar campos** – Esa misma dependencia, hace que se un modelo muy rígido, lo que complica la adición o modificación de nuevos datos para visualizar y para el modelado.

Es por ello, que a futuro se podría plantear una conexión vía API con el propio proveedor de datos. Esto otorgaría flexibilidad a la herramienta y facilitaría su uso, alineándose más con el objetivo principal de este tfm.

6.2. CLOUD RUN – CARGA DATOS

En cuanto a la implementación del sistema de carga de datos en **Cloud Run**, se ha logrado un proceso que permite la automatización parcial del flujo de información.

Una vez que el usuario ha cargado el archivo Excel con los datos de los jugadores a través de **Streamlit**, estos son enviados al sistema de **Cloud Run**, que se encarga de almacenarlos en un bucket y posteriormente trasladarlos a **BigQuery**.

Este enfoque ha mostrado ser robusto para gestionar el preprocesamiento de los datos, aunque se observan las mismas limitaciones indicadas en el apartado anterior.

6.3. MODELO

El modelo actual ha optado por un algoritmo de clasificación que, posteriormente se refina y se muestra al usuario con una estimación más cercana a una regresión.

El algoritmo utilizado muestra un nivel de precisión bastante correcto, entorno a un 82% de accuracy en el conjunto de prueba, y hasta un 95% en pruebas de validación cruzada (si bien es cierto que ese último % podría también tener un componente de overfitting).

Sin embargo, creemos que una limitación o posible mejora sería la utilización de mayor número de variables.

Actualmente el modelo se basa en ratios de evaluación de perfomance que vienen datos por el proveedor, como ELOs o RDT (Rendimiento del jugador), que son ratios que agregan un valoración del jugador sobre la cuál no se tiene constancia del proceso de cálculo. Esto puede llevar a una interpretación errónea o diferente a la que pueden tener los técnicos especialistas de fútbol del club.

Es por ello que, una posible mejora, sería obtener los datos de rendimiento de cada jugador en bruto, de forma que se pudiesen crear los estadísticos de rendimiento desde cero y poder tener un mayor control sobre qué variables incluir. De la misma forma, esto también permitiría tener mayor información sobre qué variables son relevantes en el propio proceso, pudiendo arrojar algunas conclusiones valiosas.

La disposición de mayor volumen de información, también abriría la posibilidad de poder implementar otros modelos que pudieran capturar aspectos más relevantes y ofreciesen una mayor precisión. Con un alto número de registros, se podría también plantear un modelo de regresión, e **incluso uno combinado (por ejemplo un primer modelo para clasificar en intervalos, y un segundo modelo para hacer una estimación de regresión dentro de ese intervalo)**.

6.4. CLOUD RUN – APLICACIÓN MODELO

En general, creemos que el Cloud Run del modelo es bastante completo. Nos pareció muy interesante desarrollarlo para que no tuviese un modelo fijo integrado, sino que cogiese siempre la última versión disponible en el bucket. De esta forma, el usuario no tiene que hacer cambios significativos en caso de que desee aplicar otra corrección. Bastaría con subir su nuevo modelo al bucket (siempre y cuando utilice la misma estructura de datos).

Este enfoque está muy alineado con uno de los objetivos, tal y como era la simpleza y facilidad de uso para el usuario final.

6.5. STREAMLIT

El uso de **Streamlit** ha sido uno de los puntos fuertes del proyecto, ya que ha permitido crear una interfaz amigable y fácil de usar para el usuario final. A través de esta aplicación, los usuarios pueden cargar el archivo Excel con las estadísticas de los jugadores, visualizar los datos y obtener las predicciones de precios de traspaso de manera sencilla y rápida. La interfaz no solo permite cargar el archivo, sino también realizar filtros y ordenar la información por diversos criterios, como equipo, posición o nacionalidad.

La principal ventaja de **Streamlit** es su capacidad para integrar diferentes componentes del proyecto (carga de datos, modelado y visualización) en una única interfaz, haciendo el proceso accesible incluso para usuarios con poco conocimiento técnico. Sin embargo, una posible mejora sería la optimización del rendimiento al procesar grandes volúmenes de datos, así como la integración de visualizaciones más avanzadas para facilitar el análisis de los resultados.

6.6. CONCLUSIONES GENERALES

En retrospectiva, y tras haber analizado cada una de las piezas y áreas del proyecto, se podría decir que se ha cumplido con los objetivos generales y secundarios planteados al inicio.

Estos son:

Objetivo principal:

- ✓ Desarrollar una herramienta que permita al usuario del club de fútbol realizar una estimación de los precios de traspaso de los futbolistas.

Objetivos secundarios:

- ✓ Que los desarrollos tengan en cuenta la facilidad de uso y simplicidad para el usuario final
- ✓ Que las soluciones y piezas se adapten a los conocimientos en el máster de Data Analytics de EDEM.

Es por ello, que la valoración del equipo de trabajo ha sido positiva a nivel general y estamos satisfechos con la composición final del trabajo. No obstante, somos críticos y creemos que aún hay mucho margen de mejora en la propia solución, pudiendo seguir optimizando muchas de sus piezas.

Es por ello que, nos gustaría ir un paso más allá, y detallar algunas de las posibles mejoras futuras que se podrían implantar.

7. PRÓXIMOS PASOS

El equipo se encuentra totalmente comprometido con el proyecto desarrollado. Prueba de ello es que se han identificado posibles puntos de mejora para la solución mostrada. Sin embargo, por limitaciones de tiempo de los propios desarrolladores, el alcance real ha tenido que ser más limitado.

No obstante, nos gustaría reflejarlas de cara a demostrar el potencial de la herramienta y la alineación de nuestro trabajo con los objetivos del tfm.

7.1. MEJORAS FUTURAS

A continuación se identifican algunas de los posibles futuros desarrollos para mejorar la propia solución, clasificados según su área de impacto:

Carga de datos:

- Conexión vía API con proveedores para evitar dependencias de ficheros y posibles errores por cambios de estructura.

UX:

- Una interfaz que permita jugar de forma más sencilla e individualizada.
- Posible integración con el proveedor de datos actual.

Modelado y Machine Learning:

- Desarrollar un modelo combinado, que primero clasifique y luego aplique una regresión.
- Ampliación de datos para mejorar el modelo. Esto permitiría crear nuestros propias ratios de performance.
- Adaptar un modelo específico por cada una de las posiciones posibles de un jugador.
- Desarrollo para re-entrenar el modelo cuando se carguen nuevos datos.

8. BIBLIOGRAFÍA Y REFERENCIAS

Anderson, C., & Sally, D. (2013). *The numbers game: Why Everything you know about soccer is wrong* (1ª ed.). Londres: Penguin Random House.

Cornejo, J. (2019). *Mecanismos para la Optimización del Rendimiento*. TFM. Madrid: LaLiga Business School.

Dunning, D. (2011). The Dunning-Kruger Effect. *Advances in Experimental Social Psychology, 44*, 247-296. doi:10.1016/B978-0-12-385522-0.00005-6

Gardner, M. (June de 1970). Mathematical Games. *Scientific American, 222*(6), 132-140.

Lewis, M. (2004). *Moneyball: The art of Winning an Unfair Game* (1ª ed.). Nueva York: W.W. Norton & Company.

Moreno, G. (02 de Agosto de 2017). ¿La burbuja del mercado de fichajes? *Neymar lleva al precio por traspaso a un récord histórico*. Obtenido de https://es.statista.com/grafico/10537/neymar-lleva-al-preciopor-traspaso-a-un-record-historico/

Valencia C.F. (2019). *Web Oficial Valencia C.F.* Obtenido de Palmarés: https://www.valenciacf.com/es/club/palmares