Escuela de Administración
Escuela de Ingeniería Ciencia y Tecnología

Maestría en Business Analytics

Estimación del valor en el mercado de los futbolistas en las ligas europeas

Presentado por:

Aaron Santiago Pedraza Cárdenas
Omar Ricardo Sánchez Castañeda

Bogotá, D.C. 17 de junio de 2023

Escuela de Administración
Escuela de Ingeniería Ciencia y Tecnología

Maestría en Business Analytics

Estimación del valor en el mercado de los futbolistas en las ligas europeas

Presentado por:

Aaron Santiago Pedraza Cárdenas
Omar Ricardo Sánchez Castañeda

Bajo la dirección de:

Juan David Martínez Gordillo

Bogotá, D.C. 17 de junio de 2023

Tabla de contenido

Agradecimientos

Quiero hacer un reconocimiento especial a mi director de proyecto empresarial Juan David Martínez Gordillo, que siempre tuve la mejor actitud y el mejor asesoramiento para desarrollar este documento de trabajo de grado, y a los profesores de la Maestría en Business Analytics por su apoyo incondicional durante el proceso de mi tesis. Estoy muy agradecido por la guía y el apoyo que me han ofrecido. Sus consejos y orientación han sido invaluables para mí. Estoy muy agradecido por su orientación, conocimientos y experiencias compartidas.

Aaron Santiago Pedraza Cárdenas

Mis agradecimientos al profesor Juan David Martínez Gordillo por toda la orientación brindada durante el desarrollo del presente documento y por su disposición para atender nuestras inquietudes. A todos los docentes de la Maestría en Business Analytics por los conocimientos aportados durante la duración del programa y a mis compañeros por el apoyo brindado en cada clase y por su amistad.

Omar Ricardo Sánchez Castañeda

Dedicatoria

Dedicado a mi madre por ser el principal motor de inspiración y motivación para lograr todas mis metas a nivel personas y profesional, a mi abuelo Pedro por ser un gran ser humano y un impresionante concejero, al resto de mi familia que siempre me han apoyado en cualquier iniciativa propia. Dedico a mi hermano que fue la principal persona en guiarme en este camino de realizar una maestría. Por supuesto, dedicado a todas aquellas personas que han creído en mí, que me han ayudado y acompañado en el camino para alcanzar mis metas y sueños.

Aaron Santiago Pedraza Cárdenas

Dedicado a mi mamá por ser el apoyo y soporte durante toda mi vida; a mi familia por apoyarme en el logro de cada uno de mis objetivos personales y profesionales. Dedico al presente documento a la memoria de cada uno de los miembros de mi familia que ahora guían mi camino desde el cielo y me acompañan en cada una de las pruebas que me brinda la vida.

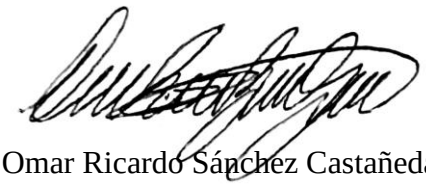
Omar Ricardo Sanchez Castañeda

Declaración de originalidad y autonomía

Declaramos bajo la gravedad del juramento, que hemos escrito el presente Proyecto Aplicado Empresarial (PAE), en la propuesta de solución a una problemática en el campo de conocimientos del programa de Maestría por mi(nuestra) propia cuenta y que, por lo tanto, su contenido es original.

Declaramos que hemos indicado clara y precisamente todas las fuentes directas e indirectas de información y que este PAE no ha sido entregado a ninguna otra institución con fines de calificación o publicación.

Aaron Santiago Pedraza Cárdenas



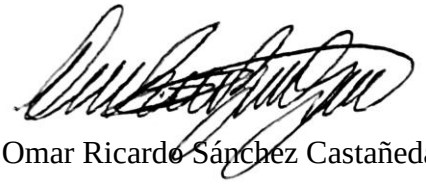
Omar Ricardo Sánchez Castañeda

Firmado en Bogotá, D.C. el 17 de junio de 2023

Declaración de exoneración de responsabilidad

Declaramos que la responsabilidad intelectual del presente trabajo es exclusivamente de sus autores. La Universidad del Rosario no se hace responsable de contenidos, opiniones o ideologías expresadas total o parcialmente en él.

Aaron Santiago Pedraza Cárdenas



Omar Ricardo Sánchez Castañeda

Firmado en Bogotá, D.C. el 17 de junio de 2023

Lista de tablas

Lista de Figuras

Glosario

1. Machine Learning: Pertenece a una ramificación de la Inteligencia Artificial encargada de convertir una serie de datos en un algoritmo competente en la extracción de patrones entre dicha serie de datos.
2. Script: Secuencia de líneas de código para conformar un programa.
3. Dataset: Conjuntos de datos almacenados en uno solo.
4. Overfitting: Consiste en que el modelo es bueno para predecir datos de entrenamiento, pero no es capaz de generalizar a otros datos.

Resumen Ejecutivo

El presente Proyecto Empresarial busca identificar las variables deportivas de los futbolistas que más relación tengan con el valor de este y así poder implementar un modelo de Machine Learning capaz de estimar el precio de los futbolistas de las diferentes ligas europeas. Lo anterior, mediante el uso del dataset (Kaggle, s. f.) en conjunto con el portal deportivo (*Transfermarkt*, s. f.).

Este Proyecto se realiza con la finalidad de poder estandarizar el cálculo del precio de los futbolistas para facilitar una estimación basada en datos deportivos. De igual manera, se busca que al ser se acceso público pueda servir como herramienta de negociación entre los diferentes clubes de fútbol.

Palabras clave

Machine Learning, Script, Transfermarkt, Kaggle, Overfitting

Abstract

This Business Project seeks to identify the sports variables of soccer players that have more relationship with its value of the same and thus be able to implement a Machine Learning model capable of estimating the price of the soccer player in the middle and end of the season of the different European leagues. The above through the use of the dataset (European Football Market Value) in conjunction with the sports portal Transfermarkt.

This Project is carried out with the purpose of being able to standardize the calculation of the price of the soccer players to facilitate an estimation based on sports data. Similarly, it is sought that being public access, it can serve as a negotiation tool between the different club's soccer.

Keywords

Machine Learning, Script, Transfermarkt, Kaggle.

1. Introducción

El presente documento aborda los contextos del mercado de futbolistas en las principales ligas europeas con el propósito de estimar el valor en el mercado de estos. Por tanto, se presenta el siguiente análisis utilizando las temáticas abordadas en la Maestría en Bussiness Analytics.

Desde inicios del siglo XXI las ligas europeas han tenido una gran relevancia en el contexto global dado el crecimiento que ha tenido el futbol a nivel mundial convirtiéndose en el deporte más practicado del planeta. Dicho aumento ha generado grandes inversiones por parte de los clubes para reforzar sus plantillas con jugadores de diferentes nacionalidades, cualidades y edades.

En la última década el mercado futbolístico se ha convertido en un referente de transacciones deportivas por los talentos de promesas del futbol o jugadores reconocidos por sus trayectorias principalmente en las ligas europeas. Esto ha generado que se generen algunos portales como (*Transfermarkt*, s. f.), especializados en la estimación del precio que puede tener un futbolista.

El origen de estos portales especializados ha permitido grandes transferencias dentro de los clubes a nivel internacional, pero ha generado condiciones de monopolio porque son estos quienes calculan los precios de los jugadores según su criterio sin que sea de conocimiento público los factores tenidos en cuenta para dicha estimación.

De igual manera, los precios difieren en cuantías significativas entre cada uno de los portales generando especulación entre los clubes que puede resultar en sobrecostos o en pérdida para equipos oferentes de jugadores. Por tanto, se hace necesario una estandarización de los criterios y publicación de la información para llegar a precios justos para cada una de las partes del negocio.

Asimismo, se presenta una problemática dado que los agentes deportivos tienden a inflar los precios de futbolistas conforme a factores como lo son la nacionalidad, la posición, la edad y el club del que proviene. Adicionalmente a esto, se tiende a que el pago de comisión de estos agentes es bastante significativo del total de la transacción dejando de lado sus estadísticas deportivas para la toma de decisiones en una venta del jugador.

Por lo tanto, el presente proyecto empresarial se plantea como la creación de un modelo predictivo mediante datos deportivos buscando estandarizar los procesos de valorización de jugadores para que una vez procesados puedan ser de consulta pública para todos los actores involucrados en el mercado del futbol como lo son clubes, entrenadores, jugadores, agentes deportivos, sponsors, periodistas y seguidores del deporte. Asimismo, el uso de esta herramienta va a servir como un mecanismo de seguimiento y evaluación al rendimiento deportivo de cada jugador.

Es importante mencionar que dicha herramienta puede ser reproducida en un entorno local siempre y cuando se cuente con la disponibilidad de datos deportivos de las ligas colombianas tanto masculinas como femeninas partiendo del hecho que la cantidad y disponibilidad de datos fue el factor decisivo para que se optara por el uso de los datos de las ligas europeas.

Por lo cual, se busca realizar un análisis descriptivo acerca de diferentes futbolistas en distintas posiciones, edades y nacionalidades para identificar los principales componentes del dataset (Kaggle, s. f.) proveniente de la plataforma Kaggle la cual es de datos abiertos.

Posteriormente, se desarrolla un análisis predictivo en busca de estimar el precio del jugador mediante sus datos deportivos.

Para este propósito, el presente documento inicia describiendo los objetivos del proyecto empresarial y su respectivo alcance. Más adelante, se establece el cronograma de cada una de las

actividades requeridas del trabajo y con ello una descripción de estas conforme a los ejes temáticos vistos en el plan de estudios de la Maestría en Bussiness Analytics.

Finalmente, se realiza una descripción del procesamiento de datos realizado como también las fases adelantadas conforme a la metodología CRISP teniendo presente una validación cruzada conforme a los resultados obtenidos.

2. Objetivos

2.1 General

Implementar un sistema que tenga la capacidad de estimar el precio de los futbolistas en el mercado mediante sus datos deportivos conforme a los datos obtenidos del dataset (Kaggle, s. f.) conectados al portal de (*Transfermarkt*, s. f.), además generar un tablero de control que permite observar las estadísticas deportivas.

2.2 Específicos

- Realizar un análisis descriptivo de las fuentes de información del dataset *(Kaggle, s. f.)* y el portal *(Transfermarkt, s. f.).*
- Creación de un visualizador de control que permita tener en cuenta las estadísticas deportivas y con ello facilidad en la toma decisiones para la venta y compra jugadores.
- Implementar una técnica de Machine Learning capaz de estimar el valor de los futbolistas mediante los datos deportivos de acuerdo a las fuentes de información donde el modelo será evaluado mediante diferentes métricas tales como R2, Mean Absolute Error, RMSE.

3. Alcance

Elaboración de un repositorio donde se encontrarán todas las fases de desarrollo del proyecto empresarial, comenzando con los diferentes notebooks para la limpieza y unión de los datos, seguido por el script de analítica descriptiva y sus diferentes fases, además de los diferentes modelos Machine Learning y sus respectivas evaluaciones. Finalmente, el acceso al tablero de control que permitirá la facilidad en la toma decisiones para venta y compra de jugadores.

4. Metodología

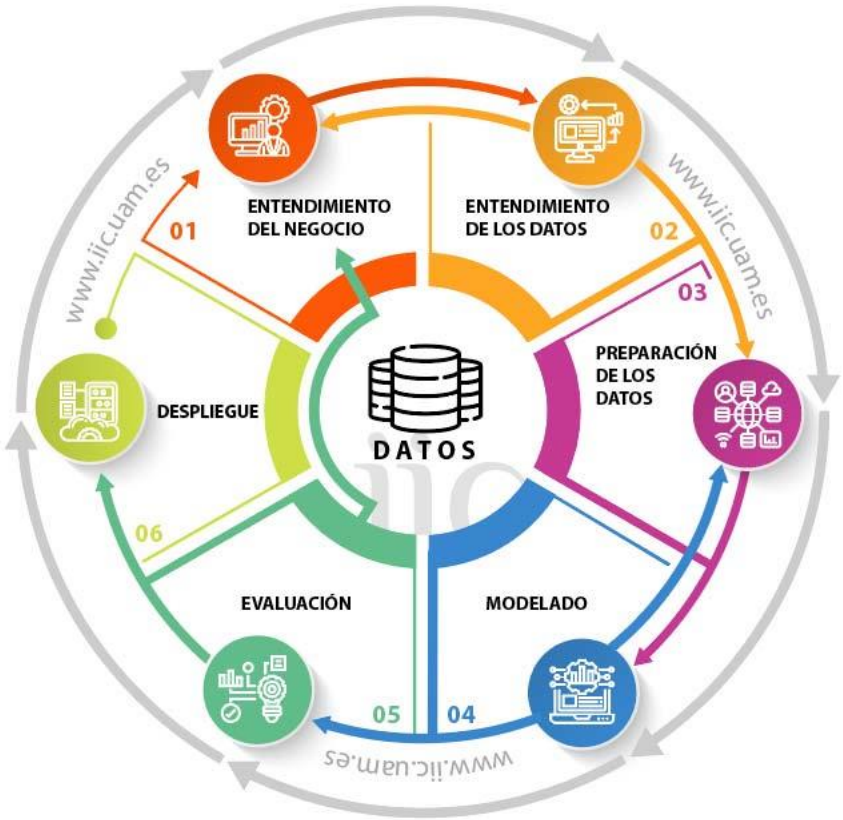
Para el presente proyecto empresarial se utilizará la metodología CRISP – DM (Gironés, Jorgi) que es un marco de trabajo versátil y eficiente utilizado para proyectos de data science. Su flexibilidad permite que se aplique en diversos contextos, en este caso, para la predicción del valor de los futbolistas.

Este enfoque proporciona una estructura coherente para el proyecto, lo cual incluye un entendimiento profundo del negocio y del valor estratégico del proyecto. Es iterativo, lo que significa que se pueden realizar modificaciones en cualquier etapa si es necesario. Al seguir un proceso estándar como CRISP-DM, se puede aumentar la eficiencia al tener una ruta clara a seguir.

En este orden de ideas, la metodología CRISP-DM se orienta mediante un entendimiento del negocio. Esto permite alinear el proyecto al objetivo estratégico, lo que en este caso sería predecir de manera precisa el valor de un jugador de fútbol.

Figura 1

Metodología CRISP



Fuente: Instituto de Ingeniería del conocimiento. Disponible en: https://www.iic.uam.es/innovacion/metodologia-crisp-dm-ciencia-de-datos/

4.1 Comprensión del Negocio

4.1.1 Determinación de objetivos del negocio

El objetivo comercial del negocio se fundamenta en la comercialización de jugadores de fútbol profesional de las ligas europeas. Principalmente, las ligas europeas están en la permanente búsqueda de talento para sus clubes para lo cual acuden a portales deportivos que utilizando datos que tienden a estimar el precio de cada uno de los jugadores.

Ahora bien, una vez los clubes obtienen un primer estimado del precio de un determinado jugador tienden a iniciar una serie de negociaciones que terminan en contratos por grandes sumas de dinero entre clubes, representantes y jugadores.

En este orden de ideas, se busca tener un desarrollo analítico de datos de libre acceso para que todos los involucrados puedan tener igualdad de acceso a la información al momento del estimativo de los precios.

4.1.2 Definir situación

En los últimos años, los mercados de jugadores de fútbol han tomado gran relevancia, debido a las altas cifras que se manejan en sus transacciones, esto ha desatado que diferentes portales deportivos como (Transfermarkt, s. f.), se especializan en la valoración del precio de los futbolistas.

Esto ha generado condiciones de monopolio dado que se estiman los precios de los jugadores según el criterio de esta plataforma (Grimaldo Santana, 2022) donde toman variables como: el nivel de la liga, la gestión del marketing de cada jugador, interés en el mercado, reputación y demás variables fuera de lo deportivo, por lo tal los valores de cada jugador no son basados en datos estadísticos deportivos.

4.1.3 Determinar objetivos de minería de datos

El objetivo de la minería de datos fundamentalmente consiste en la generación de un modelo de analítica de datos que incluya una serie de condiciones y variables que repercuten en el estimativo del costo de un jugador en el mercado de las ligas europeas. Lo anterior, buscando que

todas las partes involucradas puedan tener conocimiento de los factores que afectan el precio y partiendo de ello pueda generarse las negociaciones pertinentes.

4.2 Elaborar plan del proyecto

Tabla 1.

Cronograma de actividades

Nombre de tarea	Duración	Comienzo	Fin
Comprensión del negocio	**37 días**	**lun 11/04/22**	**mar 31/05/22**
Reunión con el Product Owner	7 días	lun 11/04/22	mar 19/04/22
Identificación de la necesidad	30 días	mié 20/04/22	mar 31/05/22
Comprensión de los datos	**75 días**	**mié 1/06/22**	**mar 13/09/22**
Concepto de los datos	15 días	mié 1/06/22	mar 21/06/22
Recolección de data	30 días	mié 22/06/22	mar 2/08/22
Exploración de datos	15 días	mié 3/08/22	mar 23/08/22
Validación de calidad de los datos	15 días	mié 24/08/22	mar 13/09/22
Preparación de los datos	**60 días**	**mié 14/09/22**	**mar 6/12/22**
Limpieza de datos	30 días	mié 14/09/22	mar 25/10/22
Análisis Descriptivo	30 días	mié 26/10/22	mar 6/12/22
Modelado	**75 días**	**mié 7/12/22**	**mar 21/03/23**
Selección de variables para el modelo	25 días	mié 7/12/22	mar 10/01/23
Selección de técnicas de modelado	25 días	mié 11/01/23	mar 14/02/23

Generación de modelo	25 días	mié 15/02/23	mar 21/03/23
Evaluación	**20 días**	**mié 22/03/23**	**mar 18/04/23**
Evaluación de resultados	10 días	mié 22/03/23	mar 4/04/23
Feedback de los modelos	10 días	mié 5/04/23	mar 18/04/23
Distribución	**30 días**	**mié 19/04/23**	**mar 30/05/23**
Creación de visualizador para resultados	15 días	mié 19/04/23	mar 9/05/23
Creación de reporte final	15 días	mié 10/05/23	mar 30/05/23

Fuente: Elaboración propia

4.3 Comprensión de los datos

4.3.1 Conceptos de los datos

La principal fuente de datos para el desarrollo del proyecto empresarial proviene de la plataforma Kaggle en específico del dataset (European Football Market Values | Kaggle, s. f.) la cual es de libre acceso al público.

El dataset de (European Football Market Values | Kaggle, s. f.) contiene información del mercado de futbolistas y datos relacionados con los jugadores de nueve ligas europeas (Premier League, La Liga, Liga NOS, Ligue 1, Bundesliga, Seria A, Premier Liga, Eredivisie y Jupiler Pro-League).

Dentro de los datos a resaltar provenientes del dataset (European Football Market Values | Kaggle, s. f.) están: Nombre del jugador, edad, posición, estatura, manager, sponsor, nacionalidad,

fecha de expiración del contrato, valor del jugador en euros y otras 15 variables deportivas asociadas al futbolista.

La segunda fuente de información proviene del portal futbolístico (*Transfermarkt*, s. f.) el cual contiene toda la información asociada al fútbol desde calendario de las ligas hasta noticias de transferencias, también como datos estadísticos deportivos de los futbolistas como partidos jugados, minutos jugados, número de tarjetas, goles, asistencias, regates, atajadas, recuperación y un sin fin de variables deportivas asociados al rendimiento de los jugadores en sus clubes y selecciones nacionales.

4.3.2 Recopilación de los datos iniciales

La primera base de datos se obtuvo mediante la página de Kaggle, la cual permite a los usuarios buscar y publicar conjuntos de datos, explorar y crear modelos en un entorno de ciencia de datos basado en la web.

Para el caso de la segunda base de datos, se toma un repositorio de GitHub (sanjitva, s. f.) el cual es open source y nos permite obtener los datos deportivos de las últimas 5 temporadas de los jugadores del portal deportivo Transfermarkt y entre las 548 variables deportivas se encuentra la variable a predecir el valor de los jugadores en euros.

4.3.3 Descripción de los datos

En esta fase se realiza una descripción de los tipos datos y número de variables que están en cada una de las bases de datos. Esta descripción se evidencia en el anexo técnico donde se realiza una breve descripción de cada variable[1].

[1] Consultar https://github.com/aaron34x/Proyect-Bussines-Analytics/tree/main/descripcion_datos

Figura 2

Descripción de variables Kaggle

```
PlayerName                          object
Affiliation                         object
League                              object
Jersey                              object
Age                                 object
birthPlace                          object
Citizenship 1                       object
Position                            object
Position 2                          object
Foot                                object
Agent                               object
ContractExpiration                  object
nationality                         object
Games Played                        object
Market Value (Euros)                object
Accumulated Transfer Sums (Euros)   object
Highest Market Value (Euros)        object
NationalTeamCaps                    object
MostRecentInjury                    object
Height (cm)                         object
dtype: object
```

Fuente: Elaboración propia

En la figura anterior, se observan las variables de la base de datos proveniente de Kaggle y los tipos de variable de cada una, como se muestra todas las variables son del tipo de dato strings (texto), por lo tanto, es importante tener en cuenta que en la fase preparación de datos se tienen que pasar las variables que tengan valores numéricos a tipo de dato int (entero).

Figura 3

Descripción de variables Transfermarkt

```
Player                                 object
Club                                   object
Age                                     int64
Position                               object
Nation                                 object
                                       ...
Own Goals (17/18)                     float64
Total Loose Balls Recovered (17/18)   float64
Aerial Duel Won (17/18)               float64
Aerial Duel Lost (17/18)              float64
% Aerial Duels Won (17/18)            float64
Length: 548, dtype: object
```

Fuente: Elaboración propia

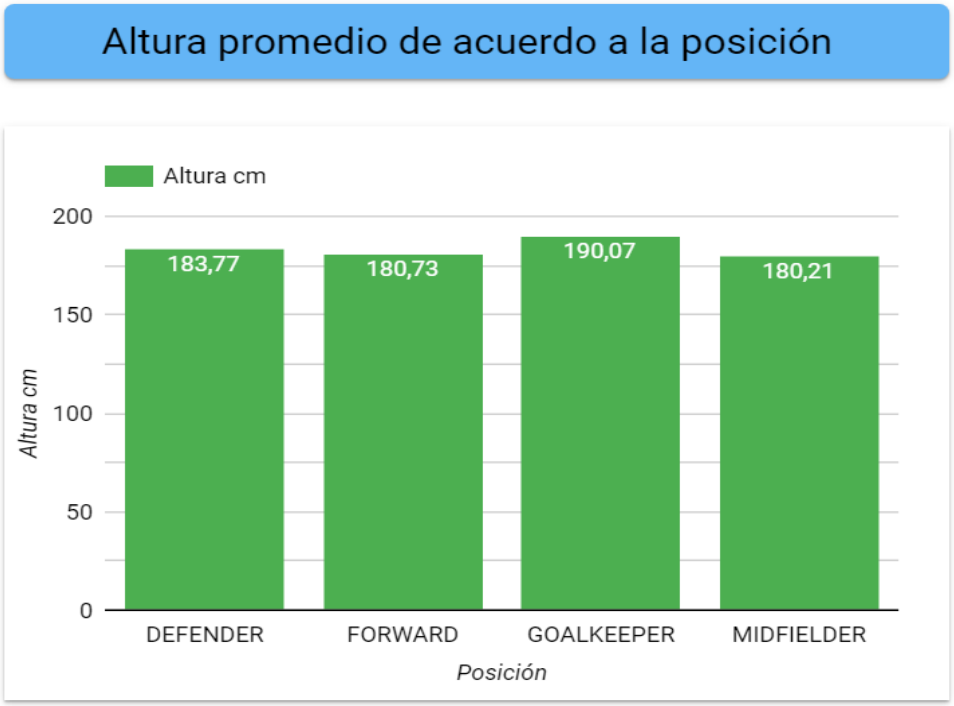
Para la segunda base de datos se toma la información proveniente del repositorio de GitHub (sanjitva, s. f) el cual se encarga de realizar un web scrapping sobre el portal de transfermarkt para así obtener un dataset con 548 variables que hacen referencia a todos los datos deportivos de las últimas 5 temporadas para 2075 jugadores de diferentes posiciones y ligas europeas. Así como en la figura anterior se observa los nombres de las variables y su respectivo tipo de dato.

4.3.4 Exploración de los datos

Seguidamente, se inicia con la fase que nos permite observar más de cerca los datos que se realizan en el proceso de minería, por lo cual mediante diferentes gráficas observamos los datos que vienen dentro de la base de datos.

Figura 4

Descripción de altura por posición en la cancha



Fuente: Elaboración propia

En la figura anterior, se puede observar los datos existentes en las bases de datos que relacionan la altura de los jugadores conforme a la posición en la que juega cada uno de ellos. Esto permite tener uno de los primeros criterios de estimación del precio dado que esto influye significativamente en el desempeño de sus actividades en el equipo.

De la misma manera, se sigue con los dineros manejados en cada una de las ligas dado que de ello depende la disposición en términos monetarios a pagar una suma específica por jugador.

En este orden de ideas, la figura anterior nos discrimina por cada liga los montos en los que son comercializados los jugadores dado que esto genera una disposición al pago de unas mejores ofertas por cada una. Por tanto, si en un mercado se manejan unos precios considerablemente mayores que los de otras ligas, esto debe verse reflejado en el costo de los

jugadores pertenecientes a estas ligas top permitiendo desprender de esto unos valores de referencia para iniciar negociaciones.

Figura 5

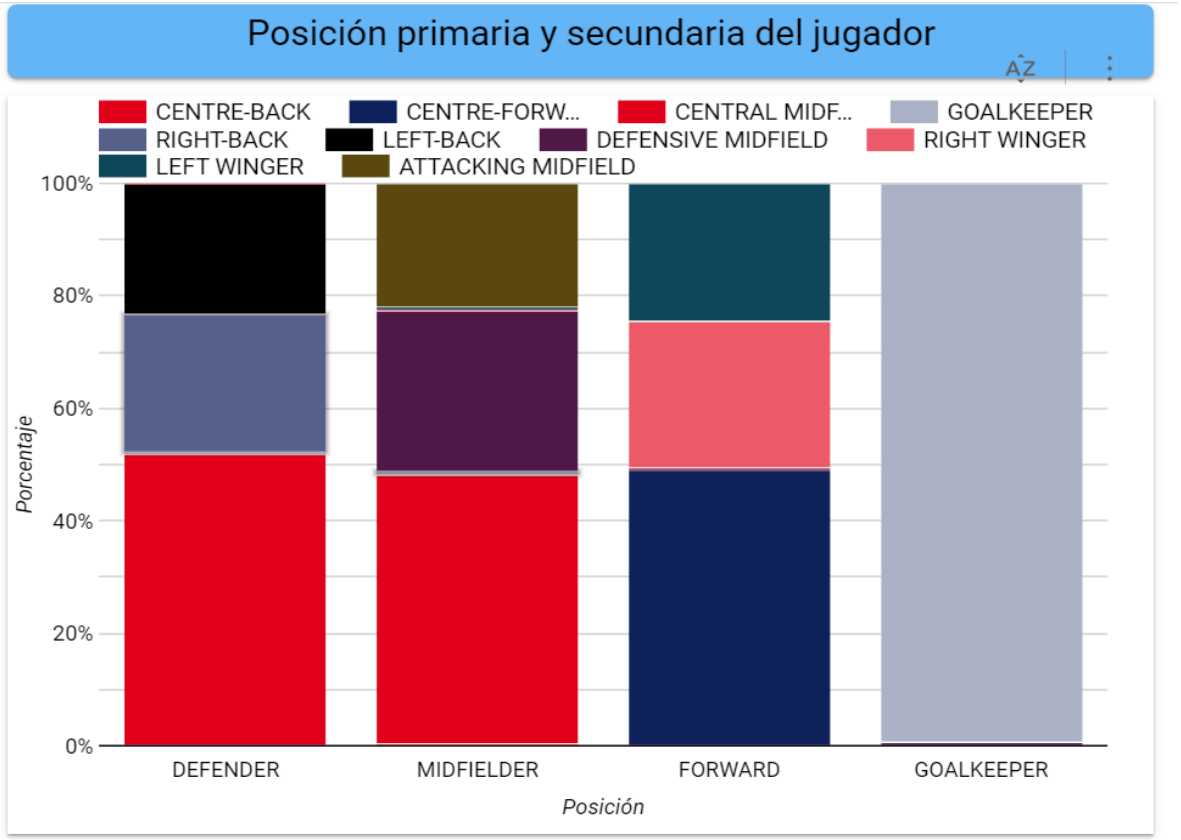
Valor promedio por liga de jugadores de las ligas europeas

Fuente: Elaboración propia

De igual importancia, tenemos la posición primaria y secundaria con el que los jugadores se ubican en el campo:

Figura 6

Relación de posición primaria y secundaria de los jugadores



Fuente: Elaboración Propia

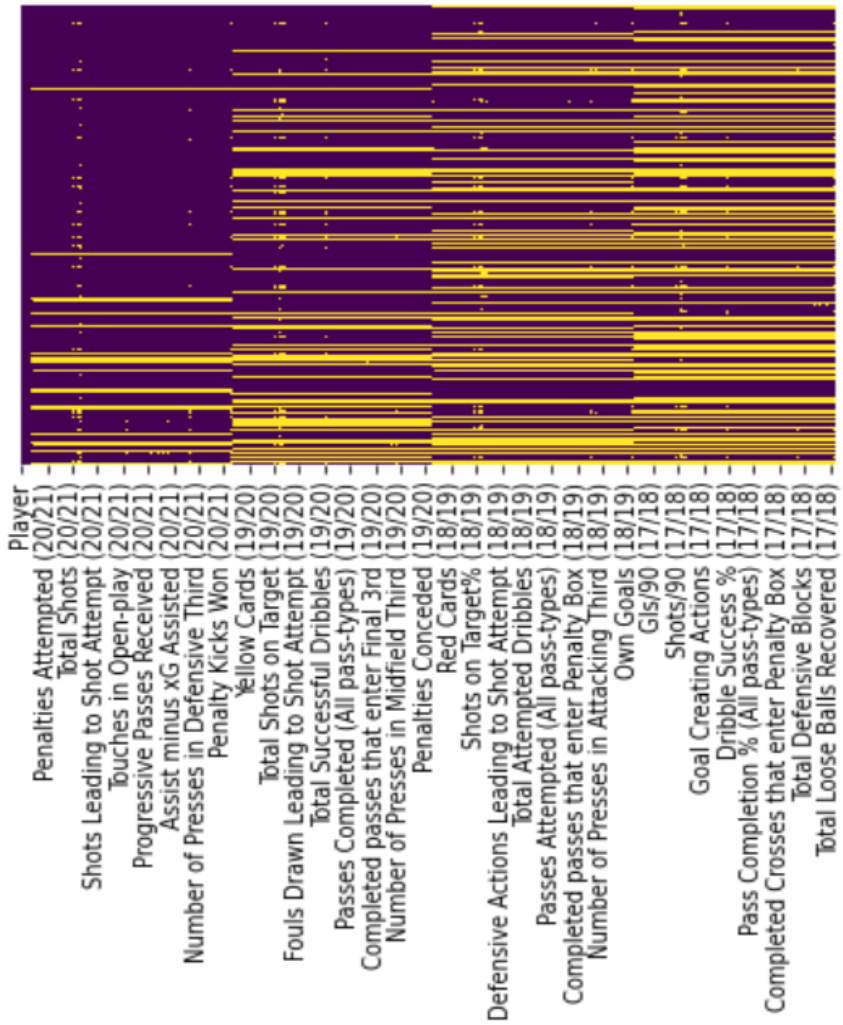
Conforme a la gráfica anterior, se puede establecer que la combinación entre las posiciones primaria y secundaria adopta una importancia significativa teniendo en cuenta que estas aumentan o disminuyen de manera significativa en el desempeño del jugador en los indicadores que su posición inicial tendería a generar altas expectativas. De igual manera, se puede resaltar que el arquero solo tiene su posición, por lo que es necesario evaluar su desempeño con indicadores únicos que no se tienen en cuenta en las posiciones de los otros jugadores.

4.3.5 Verificación de calidad de los datos

Ahora bien, una vez entendidos los datos disponibles en nuestra data podemos verificar la calidad de estos mediante métodos de verificación en Python de la siguiente manera:

Figura 7

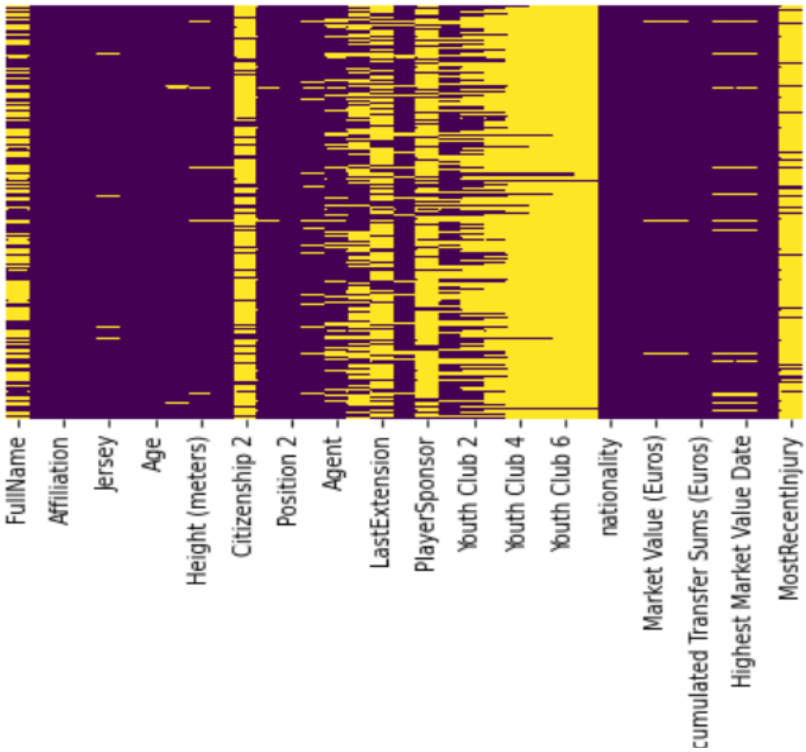
Heatmap de distribución de datos de Transfermarkt



Fuente: Elaboración propia

Figura 8

Heatmap de distribución de datos de Kaggle



Fuente: Elaboración propia

Las imágenes anteriores, nos permiten verificar que la información dispuesta en nuestras bases de datos tiene una información regular y de calidad. Esto dado que la distribución de los datos tiende a ser regular en el Heat Map se puede verificar que tiene una distribución regular.

4.3.6 Preparación de los datos

Para la preparación de los datos se realiza mediante el lenguaje de programación Python, específicamente la herramienta Google Colab.

Inicialmente, se procede a preparar y analizar cada una de las bases individualmente para detectar valores nulos, duplicados o outliers.

a) Se inicia con el procesamiento de datos de Kaggle (Kaggle, s. f.) en búsqueda de una sinergia que los homogenice para poder obtener insumos de calidad para el cálculo de los precios de mercado. Por tanto, se inicia llamando las correspondientes librerías y llamando los datos recolectados en un archivo .csv de la siguiente manera:

Figura 9

Llamado de librerías y data

Fuente: Elaboración propia

b) Una vez ejecutada esta orden al software se hace necesario verificar que hubiesen sido cargados correctamente la base de datos de Kaggle.

Figura 10

Visualización de datos puros de Kaggle

```
#Se visualiza los datos puros
df
```

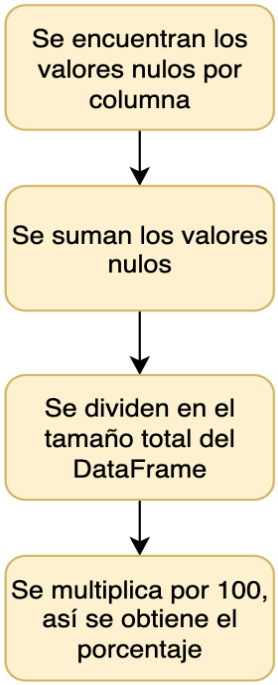
	FullName	PlayerName	Affiliation	League	Jersey	Birth Date	Age	birthPlace	Height (meters)	Citizenship 1	...	Youth Club 7	nationality	Games Played	Market Value (Euros)	U
0	Anthony Mbu Agogo Modeste	Anthony Modeste	1. FC Koln	Bundesliga	#27	4/14/1988	31	Cannes	1.87	France	...	NaN	France	1	7000000.0	
1	NaN	Benno Schmitz	1. FC Koln	Bundesliga	#2	11/17/1994	25	München	1.82	Germany	...	NaN	Germany	0	1000000.0	

Fuente: Elaboración propia

c) Ahora bien, una de las primeras técnicas abordadas consiste en la identificación de los valores nulos que se encuentran dentro de nuestra data de información. Por tanto, realiza la siguiente tarea para poder establecer la cantidad a manera de porcentaje de datos nulos que se encuentran:

Figura 11

Porcentaje de valores nulos en la base de Kaggle



Fuente: Elaboración propia

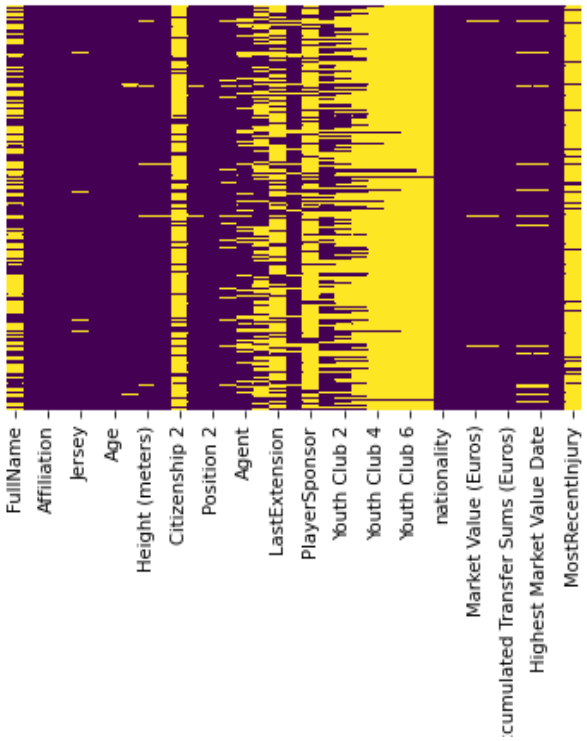
d) La figura a continuación nos brinda una idea inicial del porcentaje de valores nulos que tenemos por cada variable, sin embargo, para poder tener un mayor conocimiento se procede a realizar un Heatmap para verificar cuanto de dichos valores nulos abarca cada variable. Esto con el propósito de comprender el impacto que puede tenerse al tomar diferentes decisiones como lo pueden llegar a ser el borrar dichos valores de nuestra data:

Figura 12

Heatmap de valores nulos de Kaggle

```
# Se realiza un representacion de un heatmap de los valores nulos dentro de los Datos
sns.heatmap(df.isnull(),yticklabels=False,cbar=False,cmap='viridis')
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7fc15181e4d0>
```



Fuente: Elaboración propia

e) Este procesamiento nos brinda un primer acercamiento sobre la manera más adecuada que se debe realizar tratamiento de los datos. En este orden de ideas, se inicia tomando una de las primeras decisiones en nuestra información que consiste en eliminar las columnas que tienen más de un 30% de valores nulos. Esto nos brinda una mayor confiabilidad y credibilidad de que los valores disponibles se acercan a la realidad y especialmente no difieren de la misma por la existencia de valores extremos, que para el caso serían nulos.

Eliminar columnas con más del 30% de valores nulos es una técnica común en el preprocesamiento de datos para asegurar que la calidad y la confiabilidad de las estadísticas calculadas con estos datos sean lo más cercanas posible a la realidad[2]. Hay varias razones por las que esta decisión es justificable:

1. Mayor confiabilidad: Al preservar columnas con una menor proporción de valores nulos, se asegura que la mayoría de los datos utilizados en el análisis son válidos y representativos de la población en general.

2. Reducción del sesgo: Los valores nulos pueden introducir sesgos en el análisis si la ausencia de datos es sistemática y afecta a ciertos subgrupos de jugadores de manera desigual (por ejemplo, jugadores con pocas apariciones o jugadores que juegan en posiciones específicas).

3. Evitar errores en cálculos y modelado: Valores nulos pueden causar problemas en cálculos numéricos y procedimientos de modelado de datos. Muchos algoritmos de aprendizaje automático no pueden manejar directamente entradas con valores nulos y requerirán imputación o eliminación de registros con datos faltantes antes de construir un modelo.

4. Facilitar la interpretación: Al eliminar columnas con una alta proporción de datos faltantes, el conjunto de datos resultante será más fácil de interpretar y analizar, ya que habrá menos incertidumbre en los resultados.

Aquellas columnas que se eliminaran son 'FullName', 'Birth Date', 'Citizenship 2', 'LastExtension', 'PlayerSponsor', 'JoinedClub', 'Youth Club 1', 'Youth Club 2', 'Youth Club 3', 'Youth Club 4', 'Youth Club 5', 'Youth Club 6', 'Youth Club 7', 'Last Updated Date'.

[2] Medina Fernando, Imputación de datos: Teoría y práctica.

La eliminación de estas columnas con más del 30% de valores nulos puede considerarse justificada en este caso, ya que estas columnas podrían no ser esenciales para predecir el valor de mercado de un futbolista y debido a la alta cantidad de valores nulos, podrían no ser confiables como predictores. A continuación, se analiza cada columna y se indica por qué su eliminación podría ser apropiada:

- 'FullName': El nombre completo del jugador probablemente no influya en su valor de mercado en comparación con sus estadísticas de rendimiento y habilidades.

- 'Birth Date': La edad del jugador ya está presente en otra columna. Su fecha de nacimiento podría tener menos relevancia que su edad en términos de predecir su valor.

- 'Citizenship 2': Si un jugador tiene la doble nacionalidad, la información que aporta podría verse reflejada en otros aspectos de sus estadísticas y no tener un impacto significativo en su valor de mercado.

- 'LastExtension': La fecha de la última extensión del contrato podría no ser un factor crítico en comparación con otros factores, como los años restantes del contrato actual.

- 'PlayerSponsor': La información sobre el patrocinio del jugador podría tener menos impacto en el valor de mercado en comparación con su desempeño en el campo.

- 'JoinedClub': La fecha en que el jugador se unió al club podría tener menos relevancia en comparación con sus logros y contribuciones desde entonces.

- 'Youth Club 1' a 'Youth Club 7': Los clubes juveniles a los que asistió el jugador podrían haber influido en su desarrollo, pero su desempeño actual y sus habilidades son factores más relevantes para su valor en el mercado actual.

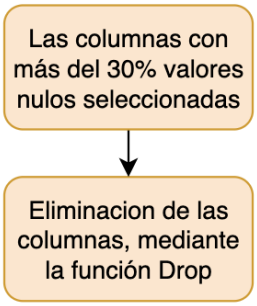
- 'Last Updated Date': La fecha en que se actualizaron los datos del jugador por última vez no debería influir en su valor de mercado.

Al eliminar estas columnas, se reduce la dimensionalidad del conjunto de datos y se enfoca el análisis en variables que están más estrechamente relacionadas con el valor de mercado de un jugador, mejorando la generalización del modelo y reduciendo el riesgo de sobreajustar a características no esenciales o no confiables.

Por tanto, imputar estos datos faltantes puede introducir más ruido a los datos, lo cual no ayudaría a mejorar el rendimiento del modelo y podría incluso empeorarlo. En este caso, eliminar estas columnas con altas proporciones de datos faltantes parece ser la opción más justificada.

Figura 13

Eliminación de columnas con valores nulos de la base de Kaggle mayor al 30%

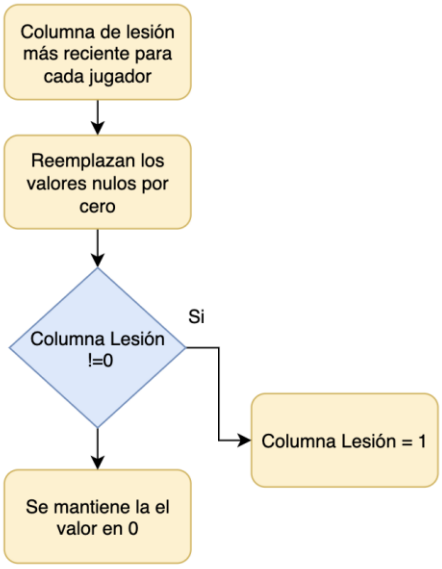


Fuente: Elaboración propia

f) La siguiente figura permite entender que la columna de lesiones recientes para jugadores la cual tiene valores nulos significando que el jugador no ha presentado lesiones, y por eso, reemplazamos esos valores nulos por cero, considerando que cero representa el hecho de que el jugador no ha tenido lesiones, mientras que uno significa que sí ha presentado alguna lesión.

Figura 14

Reemplazo de variable de lesiones de la base de Kaggle



Fuente: Elaboración propia

g) De igual manera, se identificó que los valores nulos de la columna de agente hacen referencia a que el jugador no tiene un agente asignado para su representación. En este orden de ideas, se procede a reasignar en los valores nulos en el valor de (no agent) de la siguiente manera:

Figura 15

Reasignación de agente

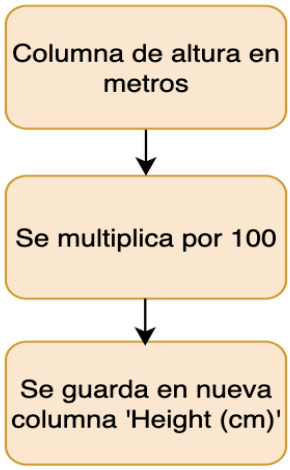
```
#Los valores nulos de la columnas de agent se reemplazan por "No agent"
# haciendo referencia que nulos es que no se conoce quien esl agente de representacion del jugador
df['Agent']=df['Agent'].fillna('no agent')
```

Fuente: Elaboración propia

h) Asimismo, para fines prácticos en la estimación del valor de los jugadores y para mejorar la precisión del modelo se realiza la conversión de unidades de metros a centímetros en lo relacionado con la altura del jugador de la siguiente manera:

Figura 16

Conversión de altura

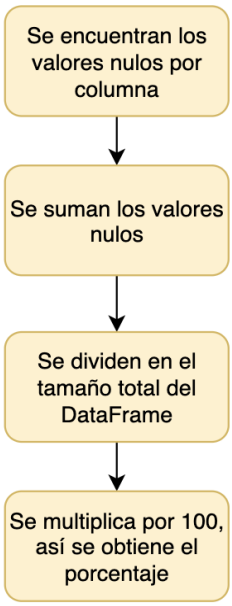


Fuente: Elaboración propia

i) Ahora bien, se procede a verificar los valores nulos una vez se ha realizado la asignación correspondiente para poder adelantar el barrido correspondiente:

Figura 17

Verificación de valores nulos



Fuente: Elaboración propia

j) Seguidamente se realiza una limpieza de los valores que se encuentren duplicados para mejorar el procesamiento de los datos:

Figura 18

Eliminación de valores duplicados

```
#en caso de que hallan valores duplicados se eliminan mediante la siguiente linea de codigo
df = df.drop_duplicates()
```

Fuente: Elaboración propia

k) De allí se puede identificar que aún tenemos algunos datos que son reconocidos como nulos por lo que se proceden a eliminar porque no se tiene información sobre ciertos jugadores ocasionando que mantenerlos en el dataset pueda introducir incertidumbre y errores en

el análisis. Por lo tanto, se dificulta hacer inferencias precisas o crear un modelo predictivo sólido cuando falta una parte significativa de la información del jugador.

Figura 18

Eliminación de valores nulos faltantes

```
# Y por ulitmo se eliminan los valores nulos de los datos que son faltantes.
df=df.dropna()
```

Fuente: Elaboración propia

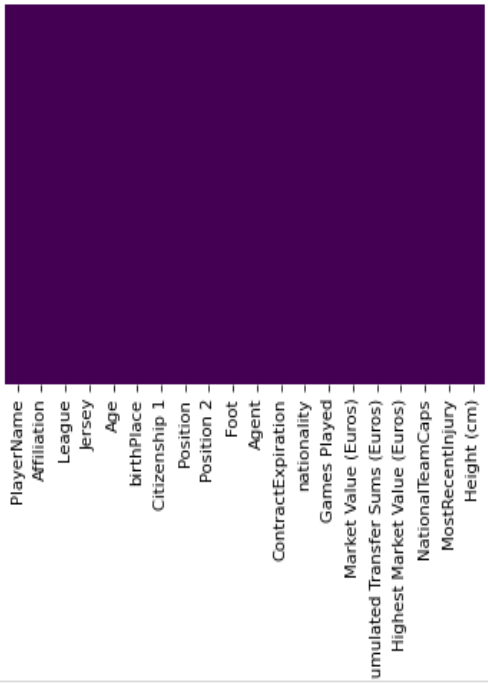
l) Esta misma verificación se puede realizar mediante un Heatmap para tener la seguridad de que los datos pueden ser procesados conforme a los requerimientos del proyecto empresarial. La uniformidad de la gráfica nos permite deducir que ya no se cuenta con valores nulos:

Figura 19

Heatmap de valores nulos de la Base de Kaggle

```
#se visualiza que de nuevo un heatmap de los valores nulos y se evidencia que nos tenemos valores nulos
sns.heatmap(df.isnull(),yticklabels=False,cbar=False,cmap='viridis')
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7fc15199e110>
```



Fuente: Elaboración propia

m) Por consiguiente, se colocan los valores que contengan texto a mayúscula para facilitar un posterior procesamiento en los datos obtenidos. Lo mismo mediante la siguiente dirección:

Figura 20

Colocación en mayúscula de valores de texto

```
#Para mayor facilidad mas adelante en el manejo de datos , se ponen en mayuscula todos valores de texto dentro de los datos
df=df.apply(lambda x: x.astype(str).str.upper())
```

Fuente: Elaboración propia

n) Uno de los primeros análisis que deben atender hace referencia a la duplicación de posición de campo del jugador pero que está señalada con un sinónimo. Por ejemplo, verificar que no existan dos posiciones como Defence y Defender que hacen referencia a la misma posición pero que están escritas de manera diferente.

Por tanto, se realiza un conteo para determinar la cantidad de valores que tenemos en dicha situación:

Figura 21

Verificación de la posición del jugador

```
# se verfica que unicamente no hallan 4 posiciones ,NOTA: ejemplo no se encuentren valores como defender y defense que hacen referencia
# a la misma Posicion pero python los toma diferentes entonces toca verificar que no hallan posicion repetidas con diferente nombre
df['Position'].value_counts()
```

```
DEFENDER      1225
MIDFIELDER    1009
FORWARD        985
GOALKEEPER     395
Name: Position, dtype: int64
```

Fuente: Elaboración propia

o) El análisis anterior se extrapola a la sección de las ligas dado que dicho inconveniente también se presenta en esta. En este orden de ideas, se procede a remitir contar los valores duplicados:

Figura 22

Verificación de posición en las ligas

```
#el mismo analisis al anterior a la linea de arriba pero para la ligas
df['League'].value_counts()
```

```
SERIE A               492
BUNDESLIGA            472
PREMIER LEAGUE        466
LALIGA                456
LIGUE 1               416
PREMIER LIGA          388
LIGA NOS              359
EREDIVISIE            352
JUPILER PRO LEAGUE    213
Name: League, dtype: int64
```

Fuente: Elaboración propia

p) Una de las primeras acciones que se adelantan para poder brindar una solución a este problema es la creación de una lista con los valores de las ligas de la siguiente manera:

Figura 23

Creación de lista con valores de ligas

```
#se crea una lista de los valores de las ligas
a=df['League'].value_counts().reset_index()
b=list(a['index'])
b
```

```
['SERIE A',
 'BUNDESLIGA',
 'PREMIER LEAGUE',
 'LALIGA',
 'LIGUE 1',
 'PREMIER LIGA',
 'LIGA NOS',
 'EREDIVISIE',
 'JUPILER PRO LEAGUE']
```

Fuente: Elaboración propia

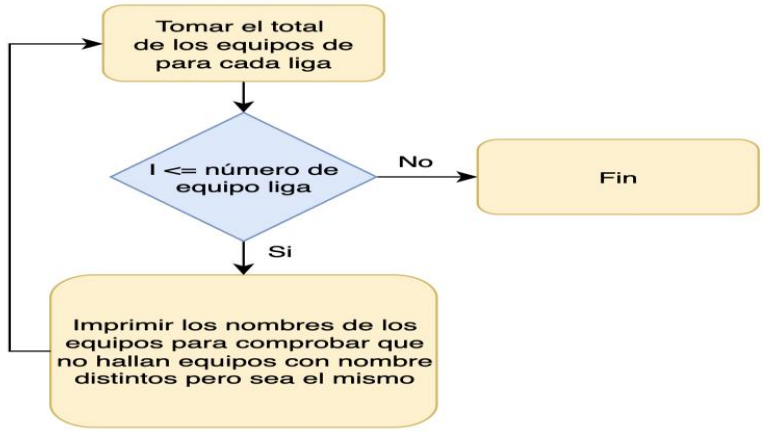
q) Seguidamente a la acción anterior, se verifica liga por liga que no se presenten equipos de diferente manera, pero hagan referencia al mismo dado que se generaría una nueva

duplicidad. Por ejemplo, R. Madrid y Real Madrid que están escritos de diferente manera, pero realmente son el mismo equipo.

Por tanto, se verifica mediante la creación una lista llamada A en la cual se guardan los valores únicos de las ligas. Posteriormente, se guardan estos valores en otra lista llamada B para realizar la siguiente acción:

Figura 24

Verificación de equipos repetidos con diferente nombre en la base de Kaggle



Fuente: Elaboración propia

La imagen anterior hace referencia a la lógica que permite verificar la duplicidad que recorre la lista B que tiene un valor único para cada liga e imprime para cada liga los equipos correspondientes permitiendo verificar que no existe duplicidad con sinónimo para cada equipo.

r) Una vez adelantado este tipo de procesamiento se verifica nuevamente el tipo de datos al que se hace relación dado que esto determinará el tipo de operación que puede hacerse con cada columna o tipo.

Figura 25

Verificación del tipo de dato

```
df.dtypes
```

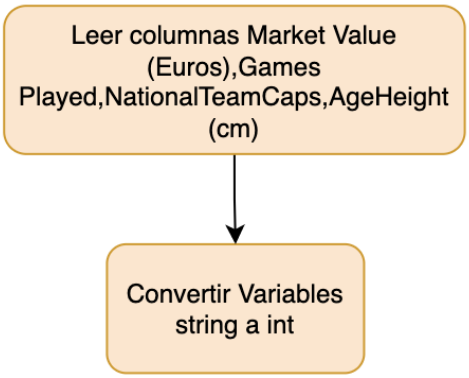
```
PlayerName                          object
Affiliation                         object
League                              object
Jersey                              object
Age                                 object
birthPlace                          object
Citizenship 1                       object
Position                            object
Position 2                          object
Foot                                object
Agent                               object
ContractExpiration                  object
nationality                         object
Games Played                        object
Market Value (Euros)                object
Accumulated Transfer Sums (Euros)   object
Highest Market Value (Euros)        object
NationalTeamCaps                    object
MostRecentInjury                    object
Height (cm)                         object
dtype: object
```

Fuente: Elaboración propia

s) Durante el proceso de análisis de la información y para fines prácticos, se toma la determinación de convertir los valores numéricos disponibles a variables de tipo entero. Lo anterior, dado el procesamiento de los datos como para facilidad de interpretación es mejor dicha presentación. La misma se realiza mediante la siguiente directriz:

Figura 26

Conversión de las columnas numéricas



Fuente: Elaboración propia

t) Finalmente, para poder verificar que la asignación fue realizada de manera adecuada dentro de nuestra base de datos de Kaggle se imprimen los datos con la asignación correspondiente:

Figura 27

Verificación de asignación de datos de Kaggle

```
df.dtypes
```

```
PlayerName                          object
Affiliation                         object
League                              object
Jersey                              object
Age                                  int64
birthPlace                          object
Citizenship 1                       object
Position                            object
Position 2                          object
Foot                                object
Agent                               object
ContractExpiration                  object
nationality                         object
Games Played                         int64
Market Value (Euros)                 int64
Accumulated Transfer Sums (Euros)   object
Highest Market Value (Euros)        object
NationalTeamCaps                     int64
MostRecentInjury                    object
Height (cm)                          int64
dtype: object
```

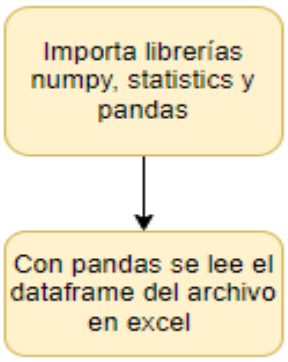
Fuente: Elaboración propia

u) Por otro lado, se realiza la preparación de la base de datos de Transfermarkt(*Transfermarkt*, s. f.) traída del repositorio de GitHub (sanjitva, s. f.) mediante un análisis similar al que se hizo con los datos de Kaggle (Kaggle, s. f.).

v) Se importan las librerías y se realiza la lectura del dataset para convertirlo en un dataframe y pueda ser más fácilmente manipulado para la limpieza.

Figura 28

Importación de librerías y lectura de dataset



Fuente: Elaboración propia

w) En la siguiente figura, se observan los porcentajes de valores nulos de la base de Transfermarkt dentro de cada variable, pero para esta base de datos no podemos eliminar estas variables ya que los valores nulos se refieren a jugadores que en esas temporadas estaban lesionados o aun no eran jugadores profesionales.

Figura 29

Porcentaje de valores nulos

```
# Se observan varias columnas tiene valores nulos,pero en este caso no podemos eliminar las columnas porque
# varios de estos valores nulos se refieren jugadores que en esas temporadas estaban lesionados o no habian debutado
# jugadores profesionales
(df.isnull().sum() / len(df))*100
```

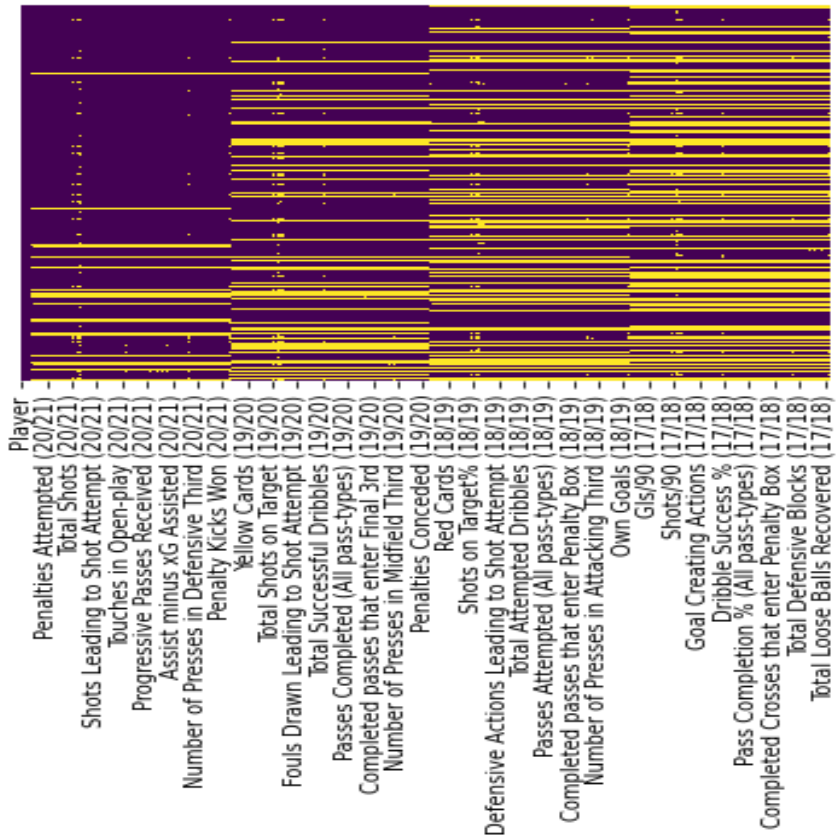
```
Player                                0.000000
Club                                  0.000000
Age                                   0.000000
Position                              0.000000
Nation                                0.000000
                                        ...
Own Goals (17/18)                    44.385542
Total Loose Balls Recovered (17/18)  44.289157
Aerial Duel Won (17/18)              44.289157
Aerial Duel Lost (17/18)             44.289157
% Aerial Duels Won (17/18)           49.301205
Length: 548, dtype: float64
```

Fuente: Elaboración propia

x) Ahora bien, mediante un Heatmap se visualiza como se distribuyen los valores nulos dentro del dataset de Transfermarkt. Lo cual evidencia que en la mayoría de las variables hay valores nulos.

Figura 30

Heatmap de dispersión de los datos nulos



Fuente: Elaboración Propia

y) Por consiguiente, como se tienen los datos deportivos de las ultimas 5 temporadas de cada jugador para los valores nulos de las diferentes variables se implementa una técnica de imputación, la cual calcula el promedio de esa variable nula con respecto a las otras temporadas y se asigna ese promedio en vez del valor nulo.

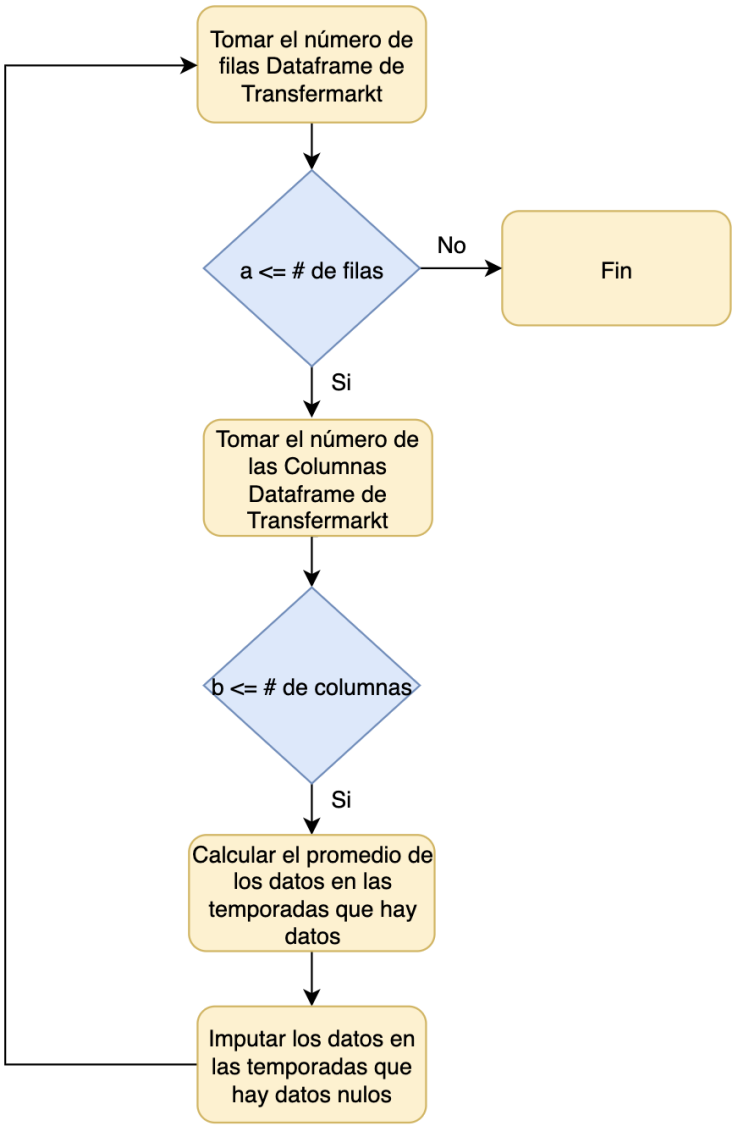
Dicha imputación suele ser útil principalmente en los deportes teniendo presente que las estadísticas deportivas tienden a ser consistentes de una temporada a otra para un jugador en particular. Por ejemplo, si un jugador tiene un alto promedio de goles en una temporada, es probable que tenga un rendimiento similar en las temporadas circundantes, a menos que haya un cambio significativo en su papel o estado físico. Entonces, el promedio de temporadas anteriores puede proporcionar una estimación razonable para los datos faltantes.

Ahora bien, esta imputación de la media a menudo puede introducir sesgo en los datos si los datos faltantes no son aleatorios. Sin embargo, al utilizar el promedio del rendimiento de un jugador a lo largo de varias temporadas, es probable que se esté proporcionando una representación más justa de su rendimiento general, reduciendo el potencial de sesgo.

Además, al aplicar esta técnica de imputación sobre los datos faltantes de esta manera, se puede mantener la integridad del conjunto de datos, sin tener que eliminar ningún jugador debido a datos deportivos faltantes. Esto permite que el modelo utilice la mayor cantidad de información posible para hacer predicciones.

Figura 31

Imputación de variables nulas



Fuente: Elaboración propia

z) Posteriormente, se realiza un análisis para poder identificar posibles valores duplicados o con nombres similares, de las siguientes variables posiciones, ligas y clubes.

Figura 32

Identificación de valores duplicados

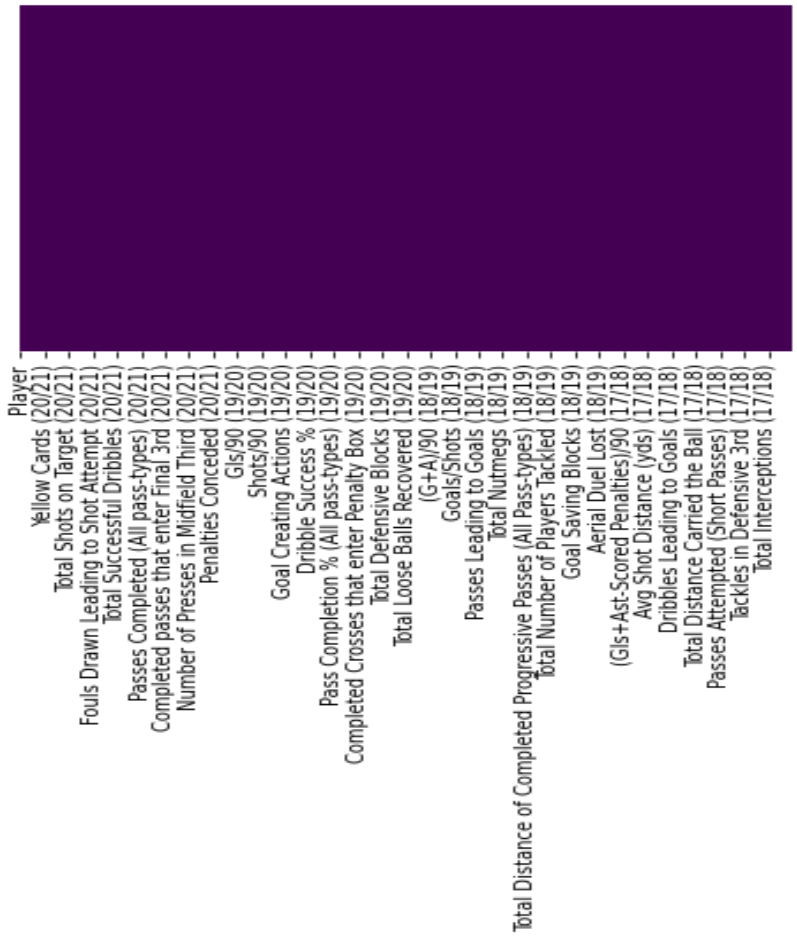
```
[ ] # se verfica que unicamente no hallan 4 posiciones ,NOTA: ejemplo no se encuentren valores como defe
    # a la misma Posicion pero python los toma diferentes entonces toca verificar que no hallan posicior
    df['Position'].value_counts()

    DEFENDER      670
    MIDFIELD      559
    ATTACK        534
    GOALKEEPER      4
    Name: Position, dtype: int64

[ ] #el mismo analisis al anterior a la linea de arriba pero para la ligas
    df['League'].value_counts()

    PREMIER LEAGUE    388
    SERIE A           378
    LA LIGA           355
    LIGUE 1           326
    BUNDESLIGA        320
    Name: League, dtype: int64

[ ] #se crea una lista de los valores de las ligas
    a=df['League'].value_counts().reset_index()
    b=list(a['index'])
    b

    ['PREMIER LEAGUE', 'SERIE A', 'LA LIGA', 'LIGUE 1', 'BUNDESLIGA']

[ ] #Se hace un for para verificar de acuerdo a cada liga que no hallan equipos repetidos con diferente
    for i in range(0,len(b)):
      a=df[df['League']==b[i]]
      print(a['Club'].value_counts())
```

Fuente: Elaboración propia

aa) Finalmente, las variables tipo texto se convierten en mayúsculas para así poder realizar más fácil la unión de los dos datasets y además se genera de nuevo un Heatmap así visualizando la dispersión de los datos nulos dentro del dataset correspondiente a los datos de Transfermarkt. Por ejemplo, se presenta dos nombres diferentes como los son R. Madrid y Real Madrid siendo que los dos hacen referencia al mismo equipo.

Figura 33

Heatmap de dispersión de datos nulos en el dataset de Transfermarkt



Fuente: Elaboración propia

La anterior figura permite verificar la homogeneidad de los datos dado que no se presentan datos nulos. Esto permite realizar el tratamiento de datos dado que con la presente distribución se encuentran aptos.

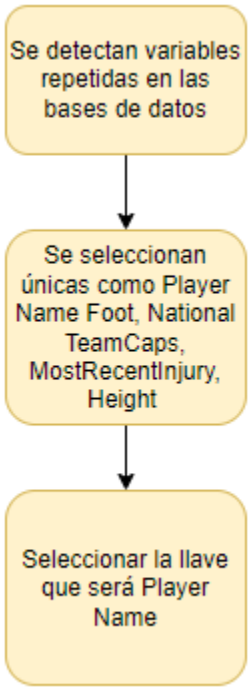
bb) En la siguiente figura, se observa el desarrollo para unir las bases de datos. Como primer paso se tienen variables repetidas dentro de las bases de datos como nacionalidad, edad y otras variables que no son útiles para determinar el valor del jugador ya establecidas con el Product Owner como número de camiseta, lugar de nacimiento, nombre del agente, entre otras. Por lo cual,

solamente se toman las variables siguientes del dataset de Kaggle: Nombre de jugador, Pie preferido, Partidos en selección nacional, si ha tenido lesiones o no, estatura en centímetros.

cc) Para el dataset de Transfermarkt se eligen todas las variables deportivas de las últimas cinco temporadas ya que esto permite conocer el rendimiento deportivo de cada futbolista en función de su posición siendo útiles más adelante para construir un modelo que estime el precio del jugador en función de sus estadísticas deportivas.

Figura 34

Selección de variables de la base de datos de Kaggle



Fuente: Elaboración propia

dd) Como se evidencia en la figura, se realiza una unión de las bases en función del nombre del jugador, para este paso es fundamental tener en cuenta que en las bases de datos los

nombre vienen sin tildes y en mayúsculas para así poder realizar un Left Join entre los datos de Transfermark y de Kaggle.

5. Modelado

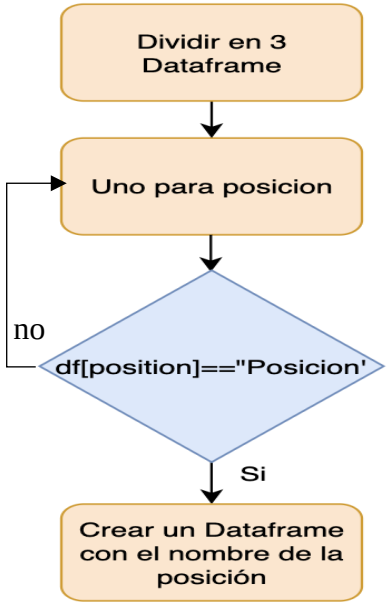
5.1 Conceptos básicos del modelado

Una vez culminadas las etapas anteriores, se procede a determinar el modelo que mejor se ajusta a nuestras necesidades y aquel que pueda garantizar un menor error con respecto al valor del mercado de referencia en el fútbol. Entonces, para el desarrollo del proyecto empresarial se hace necesario realizar un modelo regresivo teniendo en cuenta que se requiere predecir el valor de un futbolista.

Asimismo, se hace necesario la evaluación de varios modelos que se irán evaluando conforme a cada una de las tres posiciones que tienen los jugadores dado que los factores de evaluación varían entre ellos. Esto genera que deban evaluarse por separado para garantizar mayor precisión en la estimación del valor.

Figura 35

División de la posición de los jugadores en la base unida



Fuente: Elaboración propia.

Dada la necesidad que en el futbol existen cuatro principales posiciones (Arqueros, defensores, mediocampistas y delanteros) y sus estadísticas deportivas dependen de las características de cada posición; por ejemplo, los defensores tienen a no marcar los mismos goles que los delanteros, pero si tiene gran relevancia la cantidad de bloqueos y despejes en el área. Por lo tanto, es necesario generar un modelo diferente para cada posición en el campo como se realizo en la figura anterior que consiste en la creación de un dataframe para cada delanteros, mediocampistas y defensores donde no se incluyen los arqueros dados los pocos jugadores en esta posición.

5.2 Selección de técnicas de modelado

Para el presente proyecto empresarial se cuenta con diferentes técnicas disponibles que pueden estar orientadas a nuestras necesidades que son las técnicas de árbol de decisión, Modelos regresivos y Redes Neuronales.

Los modelos lineales, Random Forest y Gradient Boosting ofrecen interpretaciones claras de las relaciones de las variables. Con Random Forest y Gradient Boosting, se puede obtener la importancia de la característica que puede brindar visión sobre qué características son las más influyentes en la predicción del valor de un jugador.

Estos modelos pueden aprender y mejorar a medida que se exponen a más datos, y las Redes Neuronales pueden adaptarse a la evolución de los patrones en los datos. En este orden de ideas, se procede a evaluar cada opción de modelado para cada posición de la base unida:

- Regresión Lineal: Este modelo permite predecir el valor del precio del jugador a partir de unas variables de entrada que pueden ser altura, nacionalidad, posición secundaria, entre otras.
- Support Vector Regression (SVR): El presente modelo permite establecer una relación no lineal entre las variables de entrada de los jugadores y el precio estimado de los jugadores mediante el uso de vectores de soporte que generen la mejor regresión posible.
- Gradient Boosting Regressor: El modelo de regresión toma varios modelos de regresión de bajo nivel y mediante una serie de iteraciones puede generar un estimativo más preciso conforme a las variables de entrada de jugadores.
- Random Forest: Esta técnica de modelado permite tomar varios árboles de decisión generados con nuestras variables de entrada por jugador combinando los resultados del precio de los jugadores para encontrar un resultado lo más aproximado posible.

- Red Neuronal: Este modelo toma nuestras variables de entrada de los jugadores y mediante múltiples capas de decisión para poder estimar una predicción del precio de los futbolistas.

Lo anterior, teniendo en cuenta que deben estar orientados a que partimos de que el modelo debe ser de tipo supervisado dadas las condiciones que se tienen en el problema y objetivo empresarial.

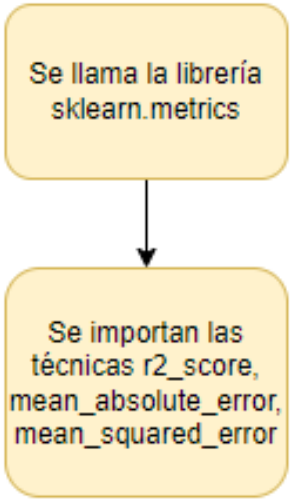
El principal requisito para aprobar un modelo es que disminuya el error lo menor posible dado que se requiere que sea fiel a las cotizaciones presentes que están en el mercado de jugadores que actualmente se tienen en las ligas europeas.

5.3 Generar un diseño de prueba

Para el proyecto empresarial buscamos tener un modelo de tipo supervisado dadas las condiciones del problema. Por lo anterior, vamos a establecer un barrido de cada modelo propuesto para el proyecto teniendo como referencia el menor valor obtenido en cada uno de los estimativos de error mediante las técnicas de r^2, Mean Absolute Error y error medio cuadrático.

Figura 36

Importe de librerías para generación de pruebas



Fuente: Elaboración Propia

5.4 Generación de los modelos

Para esta etapa del proyecto empresarial se generaron una serie de modelos regresivos y de redes neuronales que tuvieron un margen de valoración conforme a la afectación que generen sobre el error. En este orden de ideas, se tiene que cada uno de los modelos, teniendo en cuenta sus características como lo pueden ser las medidas de distribución en el caso de los modelos regresivos o el número de capas en el caso de las redes neuronales, puedan tener un mayor índice de precisión, fidelidad y mínimo error.

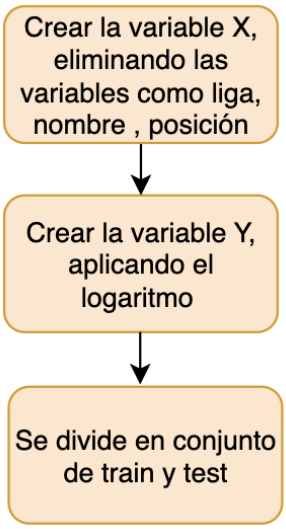
Se debe tener en cuenta que si bien ya fue realizado un proceso de tratamiento de datos donde se hizo una serie de metodologías de limpieza de datos también se realizaron unos estimativos de los no disponibles para jugadores jóvenes que tiene poca experiencia en

determinadas ligas. Esto genera que se deban establecer una serie de técnicas que eviten en el modelo una imputación agresiva.

De igual manera, se debe considerar que para aplicar nuestras técnicas de estimativo de errores primeramente fue necesario eliminar aquellas variables que fueron determinadas como irrelevantes. Para la variable Y le fue aplicado un logaritmo en base 10 con el propósito de poder manejar con una mayor facilidad el valor de los jugadores ya que son cifras grandes y esto puede llegar a afectar los modelo, y para la predicción de salida de los modelos se aplica el exponente para poder obtener el valor original en euros de los jugadores.

Figura 37

Definición de variables y aplicación del logaritmo



Fuente: Elaboración propia

Cabe aclarar que con este cambio no se altera la fiabilidad y veracidad de los resultados encontrados dado que simplemente se está realizando un tratamiento a escala para hacer manejables los datos mediante la aplicación del logaritmo.

5.5 Evaluación del modelo

Para poder seleccionar lo anterior, debemos tener en cuenta el modelo que menor error nos genere. En este orden de ideas, se procederá a evaluar cada uno de los resultados obtenidos por los modelos para poder seleccionar aquel que menor error represente al momento de brindar el valor de los futbolistas.

En este orden de ideas, podemos observar los resultados del modelo generado mediante regresión lineal:

Tabla 2

Resultados estimativos de error obtenidos por regresión lineal para defensa

TÉCNICA	R2	MEAN_ABSOLUTE_ERROR	RSME
LINEARREGRESSION	-4535828,49517191	4318063659,36365	31020751973,6378
LASSO	-0,20583688533518	9566773,03456998	15994410,6856153
ELASTICNET	-0,20583688533518	9566773,03456998	15994410,6856153
KNEIGHBORSREGRESSOR	0,25622650424125	7986813,32309097	12561581,2014411
DECISIONTREEREGRESSOR	0,20813629112551	7909615,38461538	12961318,4425517
SVR	0,36556876337863	7146566,27485578	11601555,9748923

Fuente: Elaboración propia

Tabla 3

Resultados estimativos de error obtenidos por regresión lineal para mediocampistas

TÉCNICA	R2	MEAN_ABSOLUTE_ERROR	RSME
LINEARREGRESSION	-11,2284271248457	21595238,4356553	61966018,4461473
LASSO	-0,28618836148474	12702148,45212860	20096507,0514183
ELASTICNET	-0,28618836148474	12702148,45212860	20096507,0514183
KNEIGHBORSREGRESSOR	0,05240821254497	10936309,08252370	17249596,3251263
DECISIONTREEREGRESSOR	0,19837927467295	10065154,63917520	15865463,3665408
SVR	0,26471852994412	9526998,16969177	15194804,0129614

Fuente: Elaboración propia

Tabla 4

Resultados estimativos de error obtenidos por regresión lineal para atacantes

TÉCNICA	R2	MEAN_ABSOLUTE_ERROR	RSME
LINEARREGRESSION	-6,2590750002133	26734784,4247411	63034825,0088875
LASSO	-0,26484862991126	15449815,33758440	26312322,7184807
ELASTICNET	-0,21274976961281	14945138,83498250	25764724,9279056
KNEIGHBORSREGRESSOR	0,14717171651770	11340899,12725150	21605817,6938096
DECISIONTREEREGRESSOR	0,29982502818078	12276382,97872340	19576867,4824531
SVR	0,41162927774214	9499620,48470535	17945907,9858503

Fuente: Elaboración propia

Ahora bien, también podemos observar aquellos obtenidos mediante el modelo de SVR:

Figura 38

Resultados estimativos de error obtenidos por SVR para atacantes

```
r2 0.4109239350314199
mean_absolute_error 9510301.779657718
rmse 17956661.60049359
```

Fuente: Elaboración propia

Figura 39

Resultados estimativos de error obtenidos por SVR para defensas

```
r2 0.365568763378627
mean_absolute_error 7146566.274855789
rmse 11601555.97489237
```

Fuente: Elaboración propia

Figura 40

Resultados estimativos de error obtenidos por SVR para medio campistas

```
r2 0.3699206843788748
mean_absolute_error 7153236.972234194
rmse 11561696.701249398
```

Fuente: Elaboración propia

De igual manera, podemos observar los resultados obtenidos mediante la técnica de 21:

Figura 41

Resultados estimativos de error obtenidos por Gradient Boosting Regressor para defensas

```
r2 0.6112102848608219
mean_absolute_error 5536859.772195907
rmse 9016577.289310634
```

Fuente: Elaboración propia

Figura 42

Resultados estimativos de error obtenidos por Gradient Boosting Regressor para medio campistas

```
r2 0.5085686560965526
mean_absolute_error 7680455.740794644
rmse 12519805.85676971
```

Fuente: Elaboración propia

Figura 43

Resultados estimativos de error obtenidos por Gradient Boosting Regressor para atacantes

```
r2 0.5712183199517102
mean_absolute_error 7665341.753560392
rmse 14728300.450019976
```

Fuente: Elaboración propia

Asimismo, fue estimada con la técnica de Random Forest obteniendo los siguientes resultados:

Figura 44

Resultados estimativos de error obtenidos por Random Forest para atacantes

```
r2 0.4288052432724305
mean_absolute_error 9482718.43919771
rmse 16693254.748248199
```

Fuente: Elaboración propia

Figura 45

Resultados estimativos de error obtenidos por Random Forest para medio campistas

```
r2 0.5180867711502584
mean_absolute_error 4963427.17434763
rmse 9197466.723191619
```

Fuente: Elaboración propia

Figura 46

Resultados estimativos de error obtenidos por Random Forest para defensas

```
r2 0.48543177950592475
mean_absolute_error 7565634.298704602
rmse 12415108.803849205
```

Fuente: Elaboración propia

Finalmente, se obtiene el estimativo mediante la técnica de redes neuronales arrojando los siguientes resultados:

Figura 47

Resultados estimativos de error obtenidos por Redes Neuronales para Defensores

```
r2 -0.7182139827006042
mean_absolute_error 12254942.759274215
rmse 18954963.407895207
```

Fuente: Elaboración propia

6. Evaluación

6.1 Evaluación de los resultados

Para poder evaluar y comparar los algoritmos ejecutados se aplicarán las diferentes métricas de evaluación como r^2, Mean Absolute Erro y rmse. Es por ello que conforme a los resultados obtenidos procedemos a seleccionar aquel modelo que nos brinde un menor error y que al mismo tiempo se aproxime al valor de mercado.

En este orden de ideas, el modelo Gradient Boosting Regressor resultó ser el mejor para predecir el valor de los futbolistas en euros con respecto a otros modelos, según las métricas usadas para evaluar su rendimiento.

La primera métrica utilizada para la evaluación fue R2, que es la proporción de la varianza total de la variable explicada por el modelo. En otras palabras, mide qué tan bien el modelo captura la variabilidad en los datos. Un R2 de 1 indica que el modelo explica toda la variabilidad, mientras que un R2 de 0 indica que el modelo no explica nada de la variabilidad. En este caso, el modelo Gradient Boosting Regressor para los defensores tuvo un R2 de 0.61 lo cual indica que el modelo puede explicar el 61% de la variabilidad de la variable dependiente (el valor de los futbolistas),

mucho más alto que cualquier otro modelo (el segundo más alto fue el Random Forest con un R2 de 0.48). Esta tendencia también sucede con las otras posiciones afirmando que dicho modelo es el que mejor representa la variabilidad de los datos.

La segunda métrica fue el error absoluto medio (mean_absolute_error), que mide la cantidad promedio que las predicciones del modelo se desvían de los valores reales. En otras palabras, es el promedio de la magnitud de los errores en un conjunto de pronósticos, sin considerar su dirección. Cuanto más cercano a 0, mejor es el modelo. Aquí, el modelo Gradient Boosting Regressor tuvo un error absoluto medio de 5536859.77 de euros para los defensores, que es significativamente menor que los errores absolutos medios de los otros modelos y esta métrica de evaluación también fue menor para los atacantes y medio campistas con respecto a los otros modelos.

La tercera métrica fue la raíz del error cuadrado medio (rmse), que también es una medida de la diferencia entre los valores predichos por el modelo y los valores reales. Sin embargo, a diferencia del error absoluto medio, el rmse da más peso a los errores grandes. De nuevo, el modelo Gradient Boosting Regressor tuvo el rmse más bajo (9016577.289 de euros) para los defensores, lo que indica que es inferior a los otros modelos en términos de errores grandes.

Por lo tanto, el modelo Gradient Boosting Regressor es el mejor para predecir el valor de los futbolistas en euros porque presenta la mayor capacidad para explicar la variabilidad de los datos (como lo evidencia su alto R2 para todas las posiciones. Además, el menor error promedio en las predicciones con respecto a los valores reales (mean absolute error) y la menor discrepancia en términos de errores grandes (menor rmse). Es decir, este modelo, en comparación con los demás, tiende a hacer predicciones que están más cerca de los valores reales y es más resistente a

errores grandes, lo cual es crítico al predecir el valor monetario de un jugador de fútbol, dado que errores grandes pueden tener repercusiones financieras significativas.

6.2 Proceso de revisión

Del presente ejercicio se puede verificar que el estimativo del Gradient Boosting Regressor representa el menor error para las posiciones evaluadas en el modelo con resultados cercanos al valor de mercado del futbolista y con un porcentaje de error coherente con la totalidad de los datos.

Adicionalmente, se puede observar que los modelos lineales (Lineal Regression, Lasso y ElasticNet) tienen los KPIS de los modelos más bajos presentando hasta valores negativos en r^2. Lo cual significa que no se pueden representar los valores de los jugadores mediante valores lineales.

Por otro lado, se puede comentar que los modelos de ensamble tienen los mejores resultados en términos de errores menores dado que pueden tender a entrenarse muy bien a los datos con riesgo a generarse un poco de overfitting.

Se recomienda que para futuros ejercicios se tenga en cuenta este aspecto de overfitting para con ello calcularlo y poder adoptar medidas que reduzcan el mismo. De igual manera, se recomienda aumentar la cantidad de jugadores en las diferentes posiciones dado que como se evidencio en la posición de defensores, que es donde más se cuenta con cantidad de jugadores, se tienen los mejores modelos.

Teniendo en cuenta lo anteriormente mencionado se realiza una validación cruzada para estimar el efecto que puede tener el overfitting en el modelo. Por tanto, para cada una de las

posiciones se puede verificar que se genera una mejora en el r principalmente para la posición de mediocampistas.

Figura 48

Código de verificación mediante validación cruzada

```
from numpy import mean
from numpy import std
from sklearn.datasets import make_regression
from sklearn.ensemble import GradientBoostingRegressor
from sklearn.model_selection import cross_val_score
from sklearn.model_selection import RepeatedKFold
from matplotlib import pyplot
model = GradientBoostingRegressor()
cv = RepeatedKFold(n_splits=10, n_repeats=3, random_state=1234)
n_scores = cross_val_score(model, X_train, y_train, scoring='r2', cv=cv, n_jobs=-1, error_score='raise')
n_scores=np.expm1(abs(n_scores))
print('R2: %.3f (%.3f)' % (mean(n_scores), std(n_scores)))
```

Fuente: Elaboración propia

La figura anterior nos muestra el código implementado para realizar la validación cruzada para cada una de las posiciones cambiando únicamente la variable a tratar. Con ello se obtuvieron los siguientes resultados.

Figura 49

Validación Cruzada para Defensas

```
R2: 0.881 (0.181)
r2 0.6182287497201957
```

Fuente: Elaboración propia

Figura 50

Validación cruzada para Atacantes

```
R2: 0.910 (0.233)
r2 0.5626051113598852
```

Fuente: Elaboración propia

Figura 51

Validación cruzada para mediocampistas

```
R2: 0.637 (0.233)
r2 0.48948966946676375
```

Fuente: Elaboración propia

Para las figuras anteriores, un aumento en el valor del R2 después de la validación cruzada sugiere que el ajuste inicial del modelo no estaba sobreajustado a los datos de entrenamiento. Si el modelo hubiera tenido un sobreajuste inicial, esperaríamos que el R2 se mantuviera constante o incluso disminuyera después de la validación cruzada. Sin embargo, el hecho de que R2 haya aumentado sugiere que el modelo ha generalizado bien a nuevos datos y que la validación cruzada ha ayudado a mejorar su rendimiento al prevenir el sobreajuste.

Además, un R2 más alto después de la validación cruzada también indica una mayor consistencia en el desempeño del modelo independientemente de la partición específica de los datos que se estén utilizando para el entrenamiento y prueba, lo cual es otro indicador clave de que el modelo Gradient Boosting Regressor no está sobre ajustado para ninguna posición.

6.3 Determinación de los pasos a seguir

Una vez dados los resultados se puede comentar que para poder mejorar los resultados de los modelos se hace necesario aumentar el volumen de datos dado que el presente ejercicio fue realizado con tan solo 500 jugadores lo cual se complejiza cuando se busca entrenar los algoritmos.

7. Distribución

7.1 Conceptos básicos de la distribución

Una vez con un modelo verificado para el objetivo se procede a realizar una serie de validaciones conforme a las características de un jugador particular permitiendo que se puedan tener en tiempo real valores fieles a la realidad.

Esto va a permitir en nuestra herramienta brinde a todos los interesados un valor que sea de conocimiento público y que establezca un precio por jugador informado que tenga en cuenta cada variable y característica propias del mercado según el desempeño, condición física o geográfica.

De igual manera, el análisis utilizando el modelo Gradient Boosting Regressor para predecir el valor de los futbolistas tiene varias implicaciones significativas y efectos de gran alcance para el contexto local y los posibles usuarios de la solución.

Para los clubes y organizaciones deportivas, este modelo permite evaluar con mayor precisión el valor de los jugadores, lo que es crucial para las negociaciones de transferencia, la toma de decisiones estratégicas y la planificación financiera. También puede ser utilizado para

identificar talentos infravalorados o jugadores sobrevalorados, lo cual es esencial para optimizar la asignación de recursos.

Para los futbolistas y sus representantes, este modelo puede proporcionar una evaluación objetiva de su valor de mercado, lo cual es importante al negociar contratos y acuerdos de patrocinio.

Para los aficionados y los medios de comunicación, este modelo puede ofrecer una perspectiva adicional y datos cuantitativos para discusiones sobre el valor y rendimiento de los jugadores.

7.2 Planificación de distribución

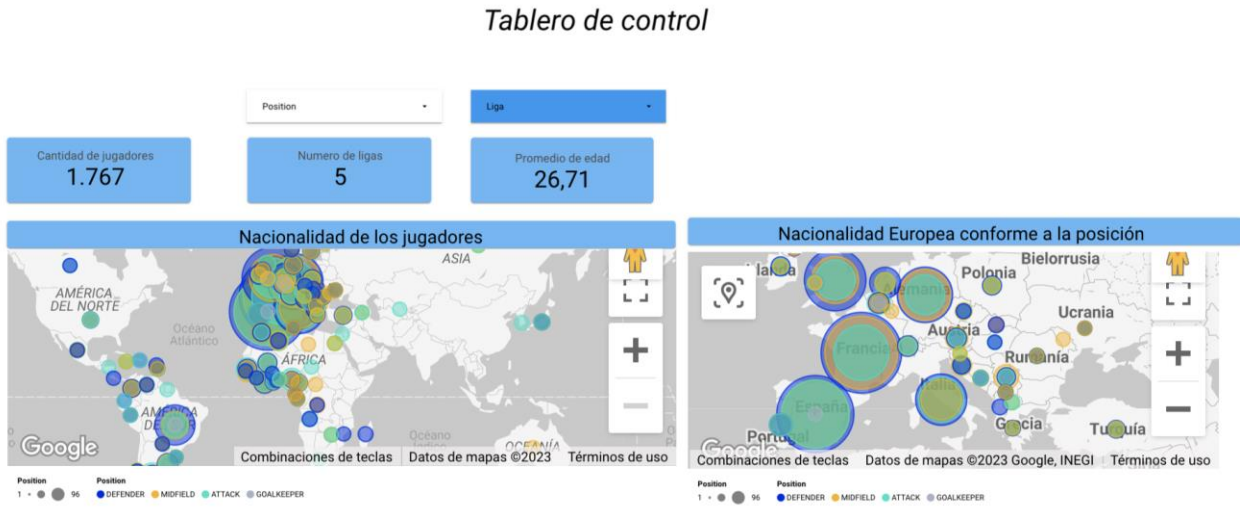
Para planificar esta distribución se busca que, mediante dos Dashboard[3] que se pueden consultar en el siguiente enlace ttps://lookerstudio.google.com/s/gIiUzUTD4Ok, el cliente pueda determinar tanto los factores que influyen en el precio del jugador como también la identificación de una serie de características que deben ser informadas a las partes mediante un tablero de control que pueden ser tenidas en cuenta en la negociación.

Los actores asociados al futbol como entrenadores deportivos, agentes, clubes, jugadores, aficionados y periodistas; pueden usar este dashboard para poder tener un control y una medición estadística sobre las métricas deportivas de los futbolistas en función de la nacionalidad, la liga, el club deportivo, entre otros.

[3] Consultar en https://lookerstudio.google.com/s/gIiUzUTD4Ok

Figura 52

Tablero de control de herramienta de estimación de valor

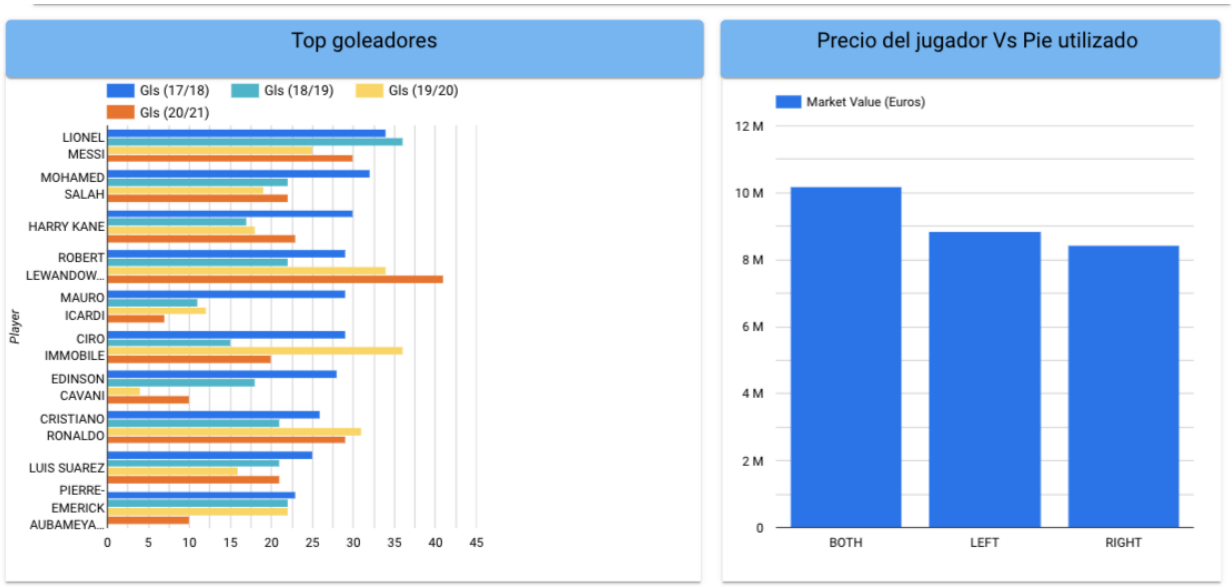


Fuente: Elaboración propia

En la figura anterior, se observa el inicio del tablero de control para la toma decisiones de los diferentes actores, siendo conformado por una serie de gráficas, kpis principales, como el número de jugadores total del dataset, el número de ligas presentes dentro del análisis y la edad promedio de los jugadores. Además, un mapa que permite analizar la distribución de jugadores en función de sus posiciones a nivel global y en específico para los países europeos.

Figura 53

Elementos Tablero de Control

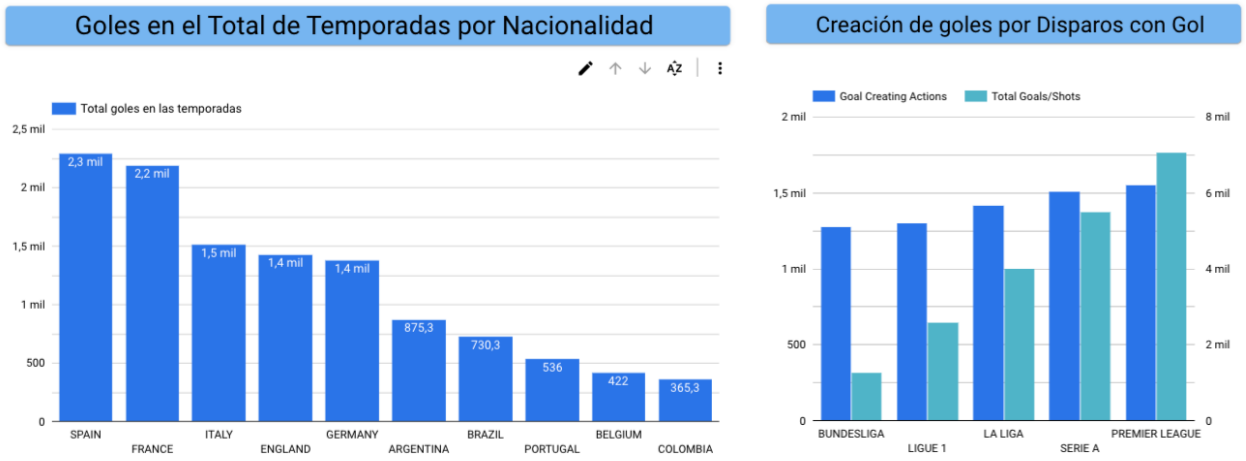


Fuente: Elaboración propia

Además, se puede observar las estadísticas de los goleadores de acuerdo a las diferentes temporadas y así poder conocer la regularidad de goles de los jugadores. También, se puede evidenciar la tendencia en la que los jugadores que manejan los dos pies tienen un mayor valor de aquellos que solo manejan un solo pie.

Figura 54

Tablero de control con data cruzada

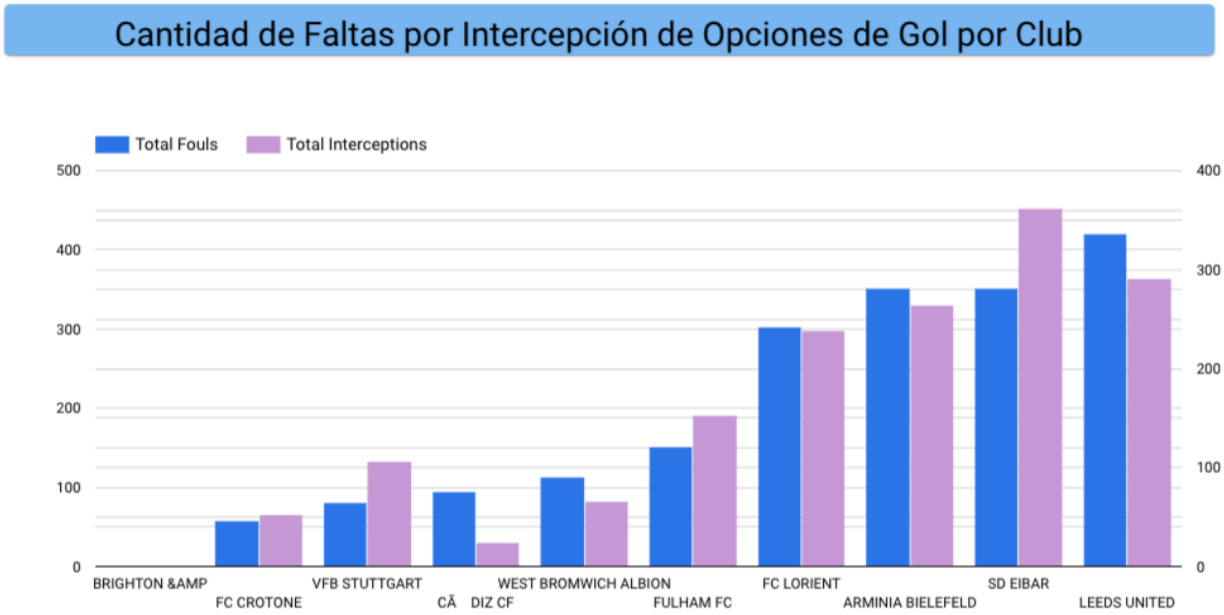


Fuente: Elaboración propia

Con respecto a la gráfica anterior, los actores asociados al futbol pueden clasificar de acuerdo con las nacionalidades cuales son los países con mayor cantidad de jugadores goleadores en la posición de delanteros y también cuales serían las ligas que más generan acciones de gol y la cantidad de goles que pueden concretarse.

Figura 55

Elementos de data cruzada para tablero de control



Fuente: Elaboración propia

Conforme a la anterior gráfica se puede inferir cuales son los equipos que más faltas comenten conforme al total de intersecciones en los partidos permitiendo a los actores deportivos conocer cuáles son las fallas que tienen al momento de interceptar balones y generar alertas en las tácticas defensivas.

7.3 Planificación del control y del mantenimiento

Dentro del tablero de control se pueden llegar analizar datos deportivos que representen un valor significativo para un actor del ámbito deportivo determinado. Un análisis como lo puede ser el hecho de que jugadores con el dominio de ambos tienen un mayor valor sobre los jugadores que solo tiene manejo de un pie. Igualmente, a este caso se pueden encontrar situaciones como la

manera en que se ven la relaciones entre entrada defensivas con las tarjetas amarrillas y rojas, como también analizar cuáles son los jugadores de la posición de delanteros, pero tiene un alto índice de fuera de lugar en función de los goles anotados.

Para poder tener un control sobre el presente proyecto empresarial se busca que se genere una actualización tanto de los datos suministrados para que se encuentren continuamente actualizados e igualmente se busca que puedan ser ingresados nuevos jugadores e información que pueda volver más robusto el modelo propuesto.

Asimismo, se deben realizar evaluaciones periódicas de preferencia al terminar cada temporada para ajustar los valores del modelo como también evitar procesos de overfitting con la información que se encuentra en la base de datos ya tratada.

7.4 Creación del informe final

Una vez finalizadas las anteriores etapas del modelo CRISP se procede a realizar un despliegue en un visualizador de datos el cual permite comparar el rendimiento de los algoritmos. Con esto se procederá a seleccionar el mejor algoritmo que será aplicado a los datos para estimar el valor de los futbolistas.

Tabla 5

Tabla de rendimiento de algoritmos

POSICIÓN	NOMBRE MODELO	R2	MEAN ABSOLUTE ERROR	MEAN SQUEARE ERROR
DEFENSAS	LINEAL REGRESSION	-11,228427148457	21595238,43560	61966018,44615
ATACANTES	LINEAL REGRESSION	-6,2590750002133	26734784,4247	63034825,00888
MEDIO CAMPISTAS	LINEAL REGRESSION	-11,228427124845	21595238,4356	61966018,44614
ATACANTES	SVR	0,411629277742135	9499620,48470535	17945907,9858503

DEFENSAS	SVR	0,365568763378627	7146566,27485578	11601555,9748923
MEDIO CAMPISTAS	SVR	0,264718529944121	9526998,16969177	15194804,0129614
DEFENSAS	GRADIENT BOOSTING REGRESSOR	0,611210284860821	5536859,7721959	0,7237401352399
MEDIO CAMPISTAS	GRADIENT BOOSTING REGRESSOR	0,445509095371827	7686636,25919977	0,7741113478351
ATACANTES	GRADIENT BOOSTING REGRESSOR	0,570514563942973	7644227,69566652	0,6635505629207
ATACANTES	RANDOM FOREST	0,442946120687110	9192577,95256898	0,8160930536864
MEDIO CAMPISTAS	RANDOM FOREST	0,496743663508239	7504389,36128435	0,7336742370159
DEFENSAS	RANDOM FOREST	0,492260698441061	5046979,86984754	0,6518926340210
DEFENSAS	REDES NEURONALES	0,00000	12249888,63803330	0,00000

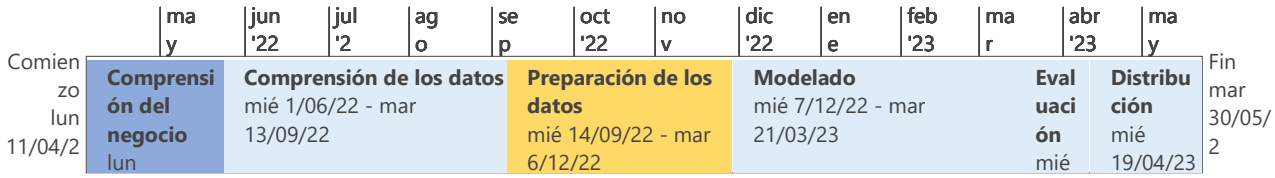
Fuente: Elaboración Propia

8. Cronograma

El siguiente cronograma presenta las correspondientes actividades a realizar en el Proyecto Empresarial teniendo en cuenta la metodología anteriormente mencionada. De igual manera, estas fechas se adelantarán conforme al desarrollo de los temas abordados en las clases de la Maestría.

Figura 56

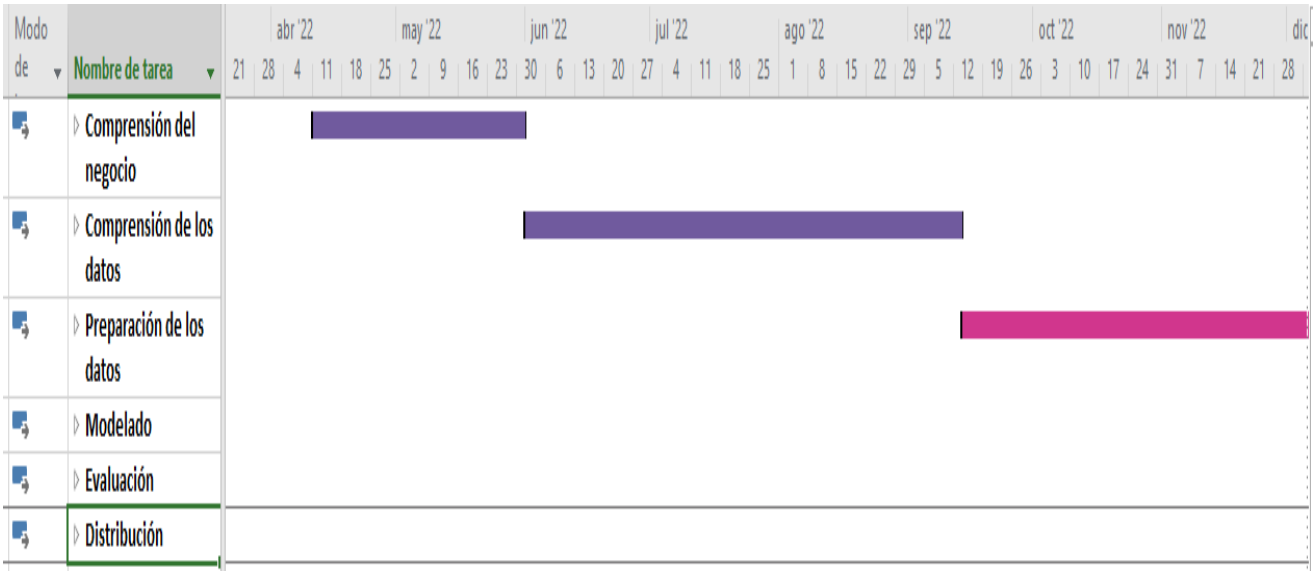
Línea de tiempo del proyecto



Fuente: Elaboración Propia

Figura 57

Hitos del proyecto empresarial

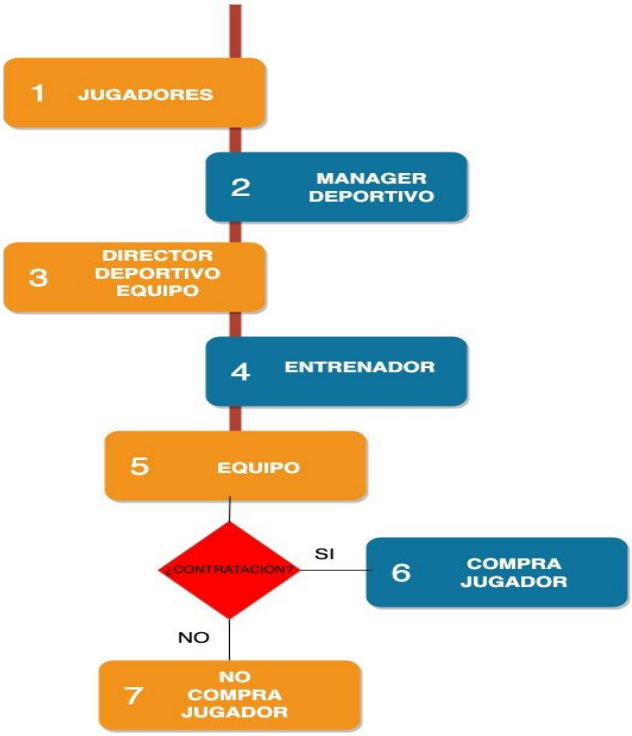


Fuente: Elaboración Propia

8.1 Descripción de la Situación organizacional donde se realizará el proyecto

Figura 58

Flujograma de decisiones actuales del mercado deportivo



Fuente: Elaboración propia

El diagrama actual del proceso inicia en el desempeño del jugador en las diferentes Ligas Europeas. Al presentarse un desempeño considerado destacable, el jugador acude a un Manager Deportivo para que resguarde sus intereses ante una eventual negociación por su fichaje entre los clubes de las ligas europeas.

El Manager Deportivo se contacta con los directores de los diferentes equipos de las ligas donde resalta las características y cualidades del jugador iniciando con ello un proceso no negociación para un posible fichaje del equipo.

Seguidamente, el director del equipo realiza la consulta con el entrenador del equipo evidenciando las posibles ventajas del fichaje y las necesidades que podrían suplirse con la

contratación de este. En consecuencia, el entrenador verifica las necesidades del equipo y coteja la información de las características del jugador con las necesidades que tiene en su nómina.

Una vez se cuenta con el análisis del entrenador, este se comunica con el equipo en su conjunto para reafirmar el interés que se tiene por realizar el fichaje. Por tanto, el equipo inicia un proceso de negociación que repercute en la contratación o no del jugador teniendo en cuenta los datos que se tienen del jugador.

8.2 Descripción de la situación estudio de caso y/o problemática empresarial y método y/o estrategia a aplicar para su solución

El problema raíz de la problemática suscitada anteriormente radica en una asimetría de información entre las partes involucradas en la negociación de futbolistas. Esta asimetría genera la comercialización entre clubes donde se beneficia aquel que cuente con los datos necesarios para determinar un precio de mercado.

Lo anterior, generando que en ocasiones pueda especularse en el valor de un deportista con valores que pueden ser superiores o inferiores a los establecidos en el mercado. Esto genera que los deportistas y clubes puedan verse afectados económicamente al recibir menos ingresos por un acuerdo o paguen un valor inflado por un jugador que no tenga un desempeño acorde con el dinero pagado.

Una evidencia presente hace un par de años consiste en el caso del futbolista Alen Halilovic (Mundo deportivo, 2022) que en el año 2014 el Futbol Club Barcelona compró a este futbolista por 2.2 millones de euros con tan solo 17 años al Dinamo Zagreb. Este jugador fue llamado por diferentes medios de comunicación como el nuevo “Messi” pero años más tarde se quedó en una

sola promesa dado que se convirtió en un jugador de una liga de bajo nivel en Europa y su valor actualmente es de 350 mil euros según el portal Transfermarkt con tan solo 26 años.

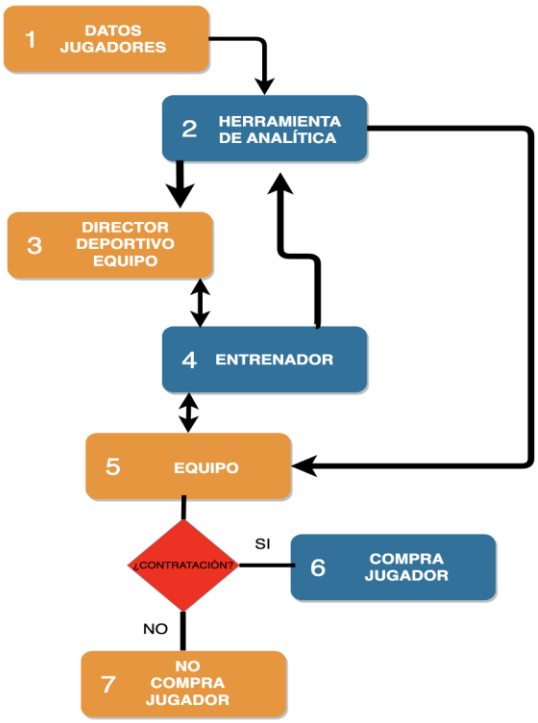
Otro caso es el exjugador Brasilero Carlos Henrique Raposo (El Espectador, 2022) quien jugo en diferentes clubes de Brasil y el mundo durante 11 años y nunca marcó un gol. Lo anterior, acudiendo a diferentes técnicas fraudulentas que ocultaban su rendimiento deportivo.

Estos casos nos permiten observar que si un jugador obtiene unos buenos datos de desempeño en algunos partidos puede generar que su precio se infle y los clubes paguen esas transferencias millonarias sin un análisis posterior que tenga en cuenta sus estadísticas deportivas a lo largo de varias temporadas.

8.3 Descripción de las alternativas, estrategias y/o acciones que se toman en el análisis de la solución a la problemática

Figura 59

Flujograma del mercado futbolístico con la herramienta de analítica



Fuente: Elaboración Propia

Con la implementación del proyecto empresarial se propone la generación de un nuevo modelo de proceso que inicia con la entrada de los datos del jugador fruto de su desempeño en las últimas temporadas de las ligas europeas.

Toda esta información alimentará nuestra herramienta de analítica que brindará información sobre las condiciones, desempeño y valor estimado del jugador para los diferentes fichajes.

Ahora bien, se tiene que el entrenador de un determinado club requiere dentro de su nómina un jugador con unas determinadas características para lo cual acude a nuestra herramienta de

analítica. En la misma puede identificar los posibles jugadores que cumplen inicialmente con los requerimientos del club.

Este entrenador se contacta con los directivos del club para comentarles las ventajas y características que tendría nominar a un determinado jugador en un club determinado. Lo anterior, apoyándose en la herramienta de analítica donde se evidenciará fácilmente el rendimiento que ha tenido el jugador en las últimas temporadas.

Teniendo en cuenta esta información, el director deportivo y el entrenador pueden iniciar una negociación basada en datos que facilite la toma de decisiones al momento de inclinarse por un determinado jugador. Con esta decisión el entrenador se puede dirigir a todas las instancias del equipo para proponer la posibilidad del fichaje.

Con todas las deducciones que se han generado entre el director deportivo y el entrenador sumado a la información que ofrece nuestra herramienta de analítica, el equipo en conjunto puede iniciar las negociaciones pertinentes para sumar a su nómina un determinado jugador.

Teniendo en cuenta toda esta información y el producto de las negociaciones el equipo puede tomar la decisión de contratar o no un determinado jugador en su plantel.

9. Caso de uso del tablero de control

9.1 Caso de uso de la situación actual del mercado futbolístico

Un equipo llamado Real Madrid se encuentra requiriendo para su equipo un defensa que pueda integrarse fácilmente a las características de su plantel como también disminuir la cantidad de ataques y goles de los equipos rivales para la próxima temporada.

En ese orden de ideas, cada club pide a sus cazatalentos consultar la disponibilidad de defensas de otros equipos que se encuentren jugando en otros equipos, ligas o en el extranjero.

En consecuencia, los cazatalentos acuden a plataformas deportivas como lo son Transfermarkt para fichar aquellos delanteros que se encuentran en tendencia por sus resultados en la última temporada como también el grado de reputación que tiene dentro de la liga.

Una vez seleccionados un par de candidatos, los cazatalentos se ponen en contacto con el jugador para hacerle una oferta y, en la mayoría de los casos, este los redirecciona con su mánager deportivo.

El mánager realiza un estimativo de la situación actual de su representado y acude las directivas del equipo en el cual se encuentra el jugador para consultar la posibilidad del traslado de este al Real Madrid. Seguidamente, estas directivas se ponen en contacto con el entrenador del equipo para evaluar el impacto que tendría esta baja para el plantel como también para obtener información que les permita obtener un estimativo del valor que tendría este jugador en caso de concretarse una transferencia al Real Madrid.

Una evaluada la posibilidad de transferencia teniendo presente el precio como también el impacto que tendría en el planten se inician las negociaciones. Las directivas del equipo interesado realizan una serie de ofertas al mánager quien al mismo tiempo realiza la consulta al club actual del defensor.

Si se llega a un acuerdo, el plantel interesado realiza el pago del valor del jugador de donde se deben descontar comisión tanto de actores como el mánager, las directivas del equipo que cede al jugador y demás intermediarios. Esto genera un sobrecosto tanto para el Real Madrid como también una perdida en términos de remuneración del jugador.

9.2 Caso de uso del proyecto empresarial

Al conocer la necesidad del equipo de contratar un defensor que pueda adaptarse a su plantel actual y tener en uso la herramienta propuesta en este proyecto empresarial:

Puede iniciar la búsqueda de jugadores, donde el primer paso es acceder al tablero de control y configurar parámetros de búsqueda donde los responsables del club configuran distintos parámetros con el perfil de jugador deseado, por ejemplo, posición: defensa con manejo de ambos pies, edad: entre 20-25 años.

Al obtener resultados de la búsqueda, el tablero de control filtra los jugadores que cumplen con estos requisitos. Se analizan los precios y estadísticas del jugador: En la lista de jugadores que cumplan los criterios de búsqueda, se despliega la información de cada jugador, incluyendo su valor de mercado actual y estadísticas recientes. Suponiendo que a través del modelo predictivo encuentran a un jugador cuyo precio es accesible para el club y cuyas estadísticas son prometedoras.

Gracias al tablero, los miembros del club pueden examinar la evolución histórica del precio del jugador que están pensando contratar. Esto les puede dar una idea de si el jugador se está valorando más con el tiempo, lo cual, podría indicar un buen potencial de crecimiento para el equipo. Con toda esta información en sus manos, los responsables del club pueden tomar una decisión más fundada sobre si vale la pena hacer una oferta para este jugador o no. En caso de decidir avanzar, los líderes del club pueden entrar en negociaciones con el club actual del jugador, utilizando la información que se tiene del modelo predictivo del precio de mercado para guiar la negociación.

Esto traería ciertas ventajas en comparación con el portal de Transfermarkt dado que el modelo utiliza técnicas de machine learning para predecir el precio del jugador con base a su

rendimiento deportivo. Esto puede generar predicciones más precisas que el método tradicional de Transfermarkt porque puede estar más influenciado por el mercado. Asimismo, se permite al usuario seleccionar criterios personalizados para buscar potenciales contrataciones de acuerdo con las necesidades específicas del equipo. Además, al ser construido en función de datos y algoritmos puede adaptarse y evolucionar con el tiempo a medida que se recopilan.

10. Referencias bibliográficas

El Espectador. (2022, noviembre 3). *El "Futbolista de farsa" que en 20 años de carrera nunca jugó un partido*. ELESPECTADOR.COM. https://www.elespectador.com/deportes/el-futbolista-de-farsa-que-en-20-anos-de-carrera-nunca-jugo-un-partido/

Grimaldo Santana, M. A. (2022, noviembre 19). *Transfermarkt: ¿Cómo se calcula el valor de un jugador de fútbol?* https://www.explica.me/noticias/Transfermarkt-Como-se-calcula-el-valor-de-un-jugador-de-futbol-20221119-0001.html

Girones Rig, Jordi. *Metodologías y estándares*. Recuperado 8 de julio de 2023, Universidad Oberta de Cataluña https://openaccess.uoc.edu/bitstream/10609/71345/4/Business%20analytics_M%C3%B3dulo%203_Metodolog%C3%ADas%20y%20est%C3%A1ndares.pdf

Kaggle. (s. f.). *European Football Market Values*. Recuperado 15 de junio de 2022, de https://www.kaggle.com/datasets/aricht1995/european-football-market-values

Medina, Fernando. *Imputación de datos,* Recuperado 9 julio de 2023, Naciones Unidad CEPAL https://repositorio.cepal.org/bitstream/handle/11362/4755/1/S0700590_es.pdf

Mundo deportivo. (2022, enero 26). *¿Qué fue de... Alen Halilovic?* Mundo Deportivo. https://www.mundodeportivo.com/futbol/20220126/1001741313/que-alen-halilovic.html

sanjitva. (s. f.). *Sanjitva/Predicting-Football-Player-Transfer-Values: Flatiron School Capstone project. Trying to find out how well players' on-field performance metrics can be used to predict their transfer values.* Recuperado 6 de diciembre de 2022, de https://github.com/sanjitva/Predicting-Football-Player-Transfer-Values

Transfermarkt. (s. f.). Recuperado 15 de junio de 2022, de https://www.transfermarkt.co/

Anexos Técnicos

Anexo A. Repositorio en Github

https://github.com/aaron34x/Proyect-Bussines-Analytics

Anexo B. Tablero de Control Herramienta

https://lookerstudio.google.com/s/gIiUzUTD4Ok

Anexo C. Descripción de variables

https://github.com/aaron34x/Proyect-Bussines-Analytics/tree/main/descripcion_datos