Análisis del impacto de cambio de entrenador en ligas profesionales de fútbol y predicción de puntos y goles posterior al cambio de entrenador con modelo de Machine Learning

Titulación:
Master U. en Big Data y Ciencia de Datos

Curso académico
2021 – 2022

Alumno/a: Suavita Losada, Andrey Fernando

Cedula de Ciudadanía Colombiana.: 1012399956

Director/a de TFM: Raul Reyero Diez

Convocatoria:

Primera

13/03/2023

De:
Planeta Formación y Universidades

Índice

Índice de ilustraciones

Índice de tablas

Resumen

El cambio de entrenador y sus consecuencias en un equipo deportivo es un tema que se lleva estudiando desde hace muchos años y ha generado controversia en las juntas de administración de los clubes deportivos a la hora de tomar decisiones sobre la continuidad del entrenador, en las ligas femeninas se han hecho escasas investigaciones sobre el tema, pues existe poca información al respecto, razón por la cual los estudios realizados se han hecho utilizando datos de ligas masculinas. El entrenador es una pieza fundamental en un equipo y su cambio se da por diferentes motivos como el mal rendimiento del equipo en ese momento, comúnmente se cree que el cambio de entrenador se da con el propósito de mejorar el rendimiento del equipo de forma instantánea, ya que el cambio de entrenador generara en los jugadores y afición la esperanza del cambio y del mejoramiento, sin embargo existen teorías que también afirman que el cambio de entrenador simplemente es una fachada para aplacar a los aficionados y que en realidad no se solucionara los problemas de rendimiento que el equipo pueda tener. En este trabajo se realizaron diversos análisis a las mejores ligas masculinas y femeninas de primera división utilizando recursos estadísticos como la regresión lineal y se realizó una comparación de medias de diversas características que poseen las ligas como el promedio de goles y el número de puntos realizados a corto plazo. Utilizando la metodología CRISP-DM se realizó adicional un proceso de recolección de datos de las ligas masculinas y femeninas profesionales de fútbol para efectuar el análisis estadístico y para construir cuatro modelos de aprendizaje supervisado con el fin predecir el comportamiento que puede llegar a tener un club deportivo en cuanto a rendimiento según varias características, como el número de puntos hecho previo al cambio de entrenador o el promedio goles hechos y recibidos previo al cambio de entrenador. Posterior al análisis se determinó que existe un efecto positivo en todas las características estudiadas de los Datasets construidos a partir de la recolección de datos futbolísticos, el efecto se extiende a las ligas femeninas variando solo en la magnitud de las medias.

Palabras clave: Cambio de entrenador, Ligas masculinas y femeninas, Fútbol, Regresión Lineal, Machine Learning

ABSTRACT

The change of coach and its consequences in a sports team is a topic that has been studied for many years and has generated controversy in the boards of directors of sports clubs when making decisions about the continuity of the coach, in women's leagues there has been little research on the subject, because there is little information about it, which is why the studies have been done using data from men's leagues. The coach is a fundamental piece in a team and its change is given for different reasons such as the poor performance of the team at that time, it is commonly believed that the change of coach is given with the purpose of improving the performance of the team instantly, since the change of coach will generate in the players and fans the hope of change and improvement, However, there are theories that also claim that the change of coach is simply a facade to placate the fans and that in reality the performance problems that the team may have will not be solved. In this work, various analyses were made of the best male and female first division leagues using statistical resources such as linear regression and a comparison of averages of various characteristics that the leagues have such as the average number of goals and the number of points scored in the short term. Using the CRISP-DM methodology, a process of data collection from the men's and women's professional soccer leagues was carried out to perform the statistical analysis and to build four supervised learning models in order to predict the behavior that a sports club may have in terms of performance according to several characteristics, such as the number of points scored prior to the change of coach or the average number of goals scored and conceded prior to the change of coach. After the analysis, it was determined that there is a positive effect on all the characteristics studied in the Datasets constructed from the collection of soccer data, the effect extends to the women's leagues, varying only in the magnitude of the averages.

Keywords: Coaching change, Men's and women's leagues, Soccer, Linear Regression, Machine Learning.

1. Introducción

En el fútbol se ha tenido la creencia de que el cambio de entrenador mejora los resultados de un equipo, sobre todo si está teniendo una mala racha, esta creencia permite a los fanáticos del deporte tener esperanza en que su equipo mejorara sus resultados rápidamente, es decir que el equipo tendrá un beneficio instantáneo en su rendimiento una vez llegue un nuevo entrenador a corto plazo, este efecto se ha estudiado en las ligas masculinas donde hay controversia en los resultados obtenidos a través de los años generando así las 3 teorías del cambio de entrenador, donde la primera indica que el cambio de entrenador traerá beneficio y bienestar al rendimiento del equipo que acaba de aceptar, la segunda teoría indica que el equipo que cambia de entrenador viene de un serie de malos resultados y el cambio de entrenador solo genera más problemas al equipo que traen como consecuencia que el equipo empeore razón que llevara al equipo a cambiar de entrenador nuevamente entrando en un círculo vicioso, finalmente la última teoría indica que el cambio de entrenador únicamente ayudara a calmar a la afición y a la prensa y que en realidad el equipo no obtendrá mejora alguna con el nuevo entrenador a corto plazo.

El cambio de entrenador es un tema que puede afectar también a las ligas femeninas, sin embargo los estudios realizados se han hecho con datos de ligas masculinas de varias divisiones, esto se debe a que las ligas femeninas no eran tan reconocidas, por lo tanto no había muchos datos en comparación con las ligas masculinas; Hoy en día las ligas femeninas han comenzado a tener más impacto y atención en las personas por lo tanto hay más datos disponibles para realizar este estudio en las ligas femeninas y poder comparar resultados.

Por lo dicho, esta investigación busca acumular un gran número de datos deportivos de las mejores ligas masculinas y femeninas de primera división incluyendo ligas europeas y latinoamericanas para realizar un análisis estadístico que permita corroborar alguna de las tres leyes del cambio de entrenador en ligas femeninas y a nivel general, también se busca construir modelos de aprendizaje supervisado utilizando los mismos datos que ayuden a los clubes deportivos a predecir el rendimiento de su equipo a corto plazo una vez se realice el cambio de entrenador, este proceso de análisis y construcción de modelos se llevara a cabo en dos partes bajo el dominio de la metodología CRISP-DM.

2. Objetivos

Objetivo general

Analizar y predecir el impacto del cambio de director técnico en equipos de las mejores ligas de fútbol de primera división a corto plazo.

Objetivos específicos

1. Recolectar datos sobre las mejores ligas de futbol masculinas y femeninas como base principal de información para utilizar en el proyecto.

2. Realizar un análisis estadístico utilizando los datos recolectados con el fin de determinar patrones o tendencias que ayuden a determinar el posible efecto beneficioso que existe en los equipos deportivos de fútbol cuando se realiza el cambio de un entrenador a corto plazo.

3. Emplear diversas técnicas de Machine Learning para predecir promedio de goles y número de puntos generados a corto plazo tras el cambio de entrenador en equipos profesionales de fútbol.

3. Teorías del efecto cambio de entrenador en el fútbol

3.1. Investigaciones previas sobre el tema de cambio de entrenador en el fútbol

La temática del cambio de entrenador en un equipo deportivo ha sido estudiada desde hace ya mucho tiempo, se han formulado diversas hipótesis buscando explicar la razón por la que un entrenador puede llegar a ser reemplazado y la consecuencia inmediata que se genera en el equipo en cuanto a rendimiento deportivo. Cuando un equipo comienza a tener malos resultados por un tiempo considerable, la junta directiva del club puede tomar la decisión de cambiar el director técnico, por otro con el fin de volver a tener buenos resultados, Grusky (1963), realizo un estudio con 16 equipos de béisbol, desde 1921 a 1941 y 1951 a 1958 donde analizo el efecto del cambio de entrenador, los resultados mostraron que se producía un déficit en el rendimiento, dando así camino para que Gamson & Scotch, (1964) realizaran un análisis de los resultados obtenidos por Grusky y pudieran elaborar las 3 teorías del reemplazo del entrenador.

La primera teoría del reemplazo de entrenador también es llamada la teoría de la casualidad unidireccional del sentido común (The common-sense one-way casuality theory) donde se afirma que el rendimiento de un equipo depende de las decisiones del entrenador, si el equipo presenta malos resultados, el entrenador es el responsable y una serie de malos resultados llevan a que el entrenador sea reemplazado con esperanza de que el nuevo entrenador presente un mejor desempeño (Peñas-Lago, 2011).

La segunda teoría conocida como la teoría de la casualidad bidireccional de Grusky (The Grusky two-way causality theory), también conocida como la teoría del círculo vicioso pues esta teoría afirma que el cambio en entrenador en un equipo repercute en el equipo pues ahora, este debe abandonar la vieja metodología del entrenador despedido y adoptar una nueva, pero este proceso conlleva tiempo y por supuesto un déficit de rendimiento. Al haber una desmejora en el rendimiento el entrenador será reemplazado de nuevo y así sucesivamente (Grusky, 1963).

La tercera teoría llamada también “la teoría del chivo expiatorio” (The ritual scapegoating no-way causality theory), sugiere que el reemplazo de un entrenador es tan solo una fachada para dar a los fanáticos del equipo nuevas esperanzas y redirigir o lanzar la culpa de la mala racha al entrenador que va a ser reemplazado (Eitzen & Yetman, 1972).

Una sucesión de malos resultados seguidos a corto plazo provoca que el entrenador tenga más probabilidades de ser despedido (d’Addona & Kind, 2012), donde los últimos partidos son los causantes del despido, y la mayoría de despidos suceden en equipos de baja tabla al borde del descenso (Audas, Dobson, & Goddard, 2002), aunque también existen equipos que pueden cambiar de entrenador si están alejados de la clasificación a la Champions League o Europa League, puede darse el caso de que algunos equipos

de media tabla cambien de entrenador para intentar clasificar a estas competiciones (Soebbing, Wicker, & Weimar, 2015), también puede darse el caso de que los entrenadores sean reemplazados si el equipo no consigue un número de puntos como es el caso de la Premier league donde el reemplazo del entrenador de un equipo puede ocurrir si este no consigue al menos 0,74 puntos por partido en promedio (Flint, Plumley, & Wilson, 2016).

Dentro de las investigaciones que se han realizado sobre la temática del efecto de cambio de entrenador, en 1997 se realizó un estudio en la Liga Inglesa, donde se analizó estadísticamente la relación entre los despidos de los entrenadores y el rendimiento de los equipos deportivos separando el ratio de victoria del 1 al 6 partido, luego del 7 al 12 partido y finalmente del 13 al 18 partido antes y después de cambiar el entrenador (Audas, Dobson, & Goddard, 1997), encontrando que las divisiones menores poseen mayor probabilidad de despido de entrenador que las divisiones mayores, y detecto una pequeña mejora en los equipos que decaía lentamente con el tiempo, en 1999 se realizó un análisis sobre el riesgo de reemplazar un entrenador teniendo en cuenta la posición en la liga el rendimiento del equipo a corto plazo, para este caso se tuvieron en encuentra los cambios voluntarios e involuntarios. En este análisis encontró que los 9 partidos antes del cambio involuntario son muy significativos para tomar la decisión, y más aún cuando la posición en la liga es baja, también apoya “la teoría del chivo expiatorio”, en donde el despido no se realiza buscando mejorar sino aplacar los medios hostiles y fanáticos enfurecidos, Audas et al. (2002) realizo otro estudio de la Liga Inglesa con datos desde 1972 hasta el 2000, utilizando un modelo de regresión Probit donde dejo de lado la teoría del chivo expiatorio y se inclinó por afirmar que el cambio de entrenador resulta incluso perjudicial dado que el equipo requiere de al menos 16 juegos para acoplarse al nuevo entrenador.

Besters, van Ours, & van Tuijl (2016) realizaron un estudio de análisis paramétrico utilizando un modelo lineal con mínimos cuadrados ordinarios (OLS) para comparar equipos que han realizado cambio de entrenador con equipos en condiciones similares que no han realizado cambio, los resultados de este análisis muestran cambios positivos al notar un aumento de rendimiento cuando se genera el cambio de director deportivo, aunque no descarta la idea de que el cambio de director deportivo pueda deberse a la “Teoría del chivo expiatorio”, en ocasiones puede ocurrir que el efecto sea negativo y haya un déficit de rendimiento en el equipo.

Heuer, Müller, Rubner, Hagemann, & Strauss (2011) realizaron un análisis estadístico de regresión en la Bundesliga temporada 1963 a 2009 diferenciando los goles de un equipo que cambió de entrenador con un equipo sin cambio, sin tener en cuenta la posición de local o visitante, resultando en que no hay diferencia aparente en los goles de un equipo tras el cambio de director deportivo a largo plazo, es decir que los aumentos de rendimientos tienen el efecto de regresión a la media.

Analizar el efecto de cambio de entrenador ha presentado un problema conocido como “regresión a la media” (A. Nevill, R. Holder, G. Atkinson, J. Copas, 2004), el cual explica que después de algunas fechas de que un entrenador nuevo ha llegado a un equipo, el efecto positivo o negativo que este haya podido tener se disipa, dando a entender que el cambio de entrenador resulta ser nulo (Cannella, A. A. & Rowe, W. G, 1995)

Hentschel, Muehlheusser, & Sliwka, (2012), aplicaron varios modelos de regresión sobre un grupo de datos de la Bundesliga con 114 reemplazos, se concluyó que los equipos deportivos aumentan su rendimiento si son homogéneos, aunque el efecto aumentado merma con el tiempo, el efecto de mejoría puede ser mayor en jugadores que tenían menor rendimiento, apoyando así la primera teoría de cambio de entrenador (The common-sense one-way casuality theory).

Soebbing et al. (2015), realizaron un estudio para la Bundesliga desde la temporada 2000 a la temporada 2013 aplicando 2 modelos de regresión Logit y 2 modelos de regresión de mínimos cuadrados ordinarios, donde obtuvo como resultado la premisa de que hay cambios positivos en el equipo luego del cambio de director deportivo, aunque este efecto se tarda aproximadamente 8 semanas en ocurrir mientras el equipo se acopla al nuevo modelo de juego propuesto por el nuevo entrenador.

Wagner (2010), realizo un modelo de regresión multivariado para la Bundesliga primera división teniendo en cuenta 4 juegos antes y después del cambio de entrenador, los resultados apoyan la teoría del sentido común (The common-sense one-way casuality theory), afirmando que los cambios de entrenadores generan un efecto positivo en el equipo sobre todo al jugar de local.

Peñas-Lago (2011), realizó un análisis basado en un modelo de regresión lineal, en una serie de juegos antes y después del reemplazo de un entrenador en un club deportivo (juegos 1,2,3,5,10,15 y 20) donde encontró que hay una mejora en el equipo, pero la ratio de victorias comienza a disminuir luego del 5 juego después del cambio de director deportivo, dado así una mejora a corto plazo, pero a largo plazo la mejora se disipa.

Algunos autores apoyan la teoría de la mejora de rendimiento que se produce al cambiar de entrenador, por ejemplo, Hentschel et al. (2012) realizo un ejemplo con equipos a los que primero se les midió la heterogeneidad, algunos habían cambiado de entrenador y otros no, descubriendo así que los equipos con más homogeneidad mejoraban más que los otros a corto plazo, pero este rendimiento aumentado comienza a disminuir hasta anularse a partir del cuarto partido, concluyendo que si hay un efecto positivo que merma con el tiempo hasta desaparecer, el efecto aumentado suele mejorar los resultados un 20% (Peñas-Lago, 2011), hay autores que apoyan el aumento de rendimiento de cambio de entrenador en encuentros que se realizan jugando en casa solo en los primeros dos encuentros y esto debido al aumento de ánimos provenientes de los fanáticos (Wagner, 2010).

El efecto de cambio de entrenador puede ser negativo; se afirma que un equipo requiere de por lo menos 16 encuentros para adaptarse a un nuevo entrenador lo que causa que en esos 16 juegos haya una desmejora en el equipo (Audas et al, 2002).

También hay investigaciones de autores que encuentran en una posición neutral, como por ejemplo Audas et al. (1997), o Besters et al. (2016), que afirman no haber cambio positivo o negativo frente al cambio de entrenador, ya que de sus resultados encontraron cambios con poca significancia al comparar equipos que no cambiaron de entrenador con equipos que sí, solo encontraron un aumento de goles (Heuer et al, 2011), y de estos autores se fundamenta principalmente la teoría del chivo expiatorio que

básicamente encuentra el cambio de entrenador como un movimiento que poco tiene que ver con mejorar el rendimiento del equipo, pero que a cambio calmara las enojadas aficiones y la prensa.

3.2. Análisis de investigaciones previas

Tras realizar una revisión de algunos autores enfocados en estudiar el efecto del cambio de entrenador producido en los equipos deportivos de varios deportes, se puede observar que existe controversia entre los estudios ya cada autor encontró evidencias para apoyar alguna de las tres teorías del cambio de entrenador. Se han estudiado algunas de las mejores ligas del mundo como la Premier league o la Bundesliga. Algunos autores apoyan la hipótesis de que el cambio de entrenador beneficia el rendimiento del equipo a corto plazo pero gracias al efecto de la “regresión a la media”, el equipo regresa a la normalidad, otros autores sugieren que el incremento de rendimiento tendrá lugar una vez el equipo se adapte a las estrategias del nuevo entrenador lo que dará lugar a buenos resultados a largo plazo pero también a resultados iguales o peores a los que se tenían previo al cambio de entrenador a corto plazo, finalmente algunos autores apoyan la tercer teoría del cambio de entrenador que indica que el cambio de entrenador simplemente ayudara a calmar los comentarios y ataques de la prensa y de los aficionados. En todos los anteriores casos se tuvo en cuenta datos de una misma liga únicamente, de primera y segunda división.

Las ligas femeninas no han sido estudiadas en las investigaciones de los autores vistos anteriormente, esto puede deberse a la alta popularidad de las ligas masculinas, y también la cantidad de datos disponibles que ofrecen las ligas masculinas, sin embargo, en la actualidad las ligas femeninas de futbol han comenzado a gustar a los fanáticos y a ser más relevantes que en la antigüedad por lo que es posible encontrar datos recientes referentes a marcadores y resultados para realizar estudios estadísticos que permitan incrementar el conocimiento que se tiene sobre la temática del cambio de entrenador, bajo la hipótesis de que existe un efecto positivo en los equipos femeninos que cambian de entrenador.

4. Técnicas de aprendizaje supervisado

Para la realización de este proyecto se utilizaron diversas técnicas de aprendizaje supervisado para tareas de análisis en el caso de la regresión lineal y predicción en el caso de los árboles de decisión y los conjuntos de modelos.

Aunque la regresion lineal es una tecnica de aprendizaje supervisado usada comunmente para realizar predicciones, tambien se puede utilizar para visualizar la tendencia de dos variables. Para este proyecto los coeficientes de la regresion lineal pueden ayudar con el analisis estadistico de las ligas profesionales de fútbol, en caso de que dos variables seleccionadas posean una alta coorelacion, la linealidad sera evidente. Por ejemplo las variables correspondientes a resutlados previos al cambio de entrenador y posteriores al cambio de entrenador pueden llegar a presentar linealidad.

La tecnicas de Decision Tree y los conjutos de modelos se utilizaron para realizar predicciones de las caracteristicas seleccionadas en la fase 3, ya que dada la alta dispercion de los datos, los conjuntos de modelos suelen obtener mejores resultados, que los obtenidos por la regresion lineal en prediccion de variables.

A continuacion se describen las caracteristicas y los hiperparametros de cada tecnica de aprendizaje supervisado.

- ***Regresión lineal***

La regresión lineal es una de las técnicas más básicas de aprendizaje supervisado, la cual se utiliza para predecir variables dependientes de una o varias variables independientes, si la variable dependiente está vinculada a más de una variable independiente, será regresión lineal múltiple, pero si la dependencia es hacia una sola variable independiente entonces el proceso será llamado regresión lineal simple.

En la librería de Scikit-learn se puede encontrar el modelo LinearRegression para la construcción de modelos de regresión lineal utilizando el método de los mínimos cuadrados ordinarios, los hiperparámetros para este modelo son los siguientes: (Pedregosa, 2011)

- **fit_intercept:**
  Toma en cuenta la intercepción o no para el modelo, si no se toma en cuenta, se espera que los datos estén centrados.

- **copy_X:**
  Este hiperparámetro genera una copia de X (variable independiente) para ser sobre escrita.

- **n_jobs:**

  Determina el número de cálculos a utilizar por el cálculo, lo cual genera aceleración en el cálculo en caso de un gran número de problemas.

- **positive:**
  Si se selecciona este Hiperparámetro, fuerza los coeficientes a ser positivos.

- ***RandomForest:***

El Random Forest es una técnica de aprendizaje supervisado utilizada en el machine learning para realizar procesos de clasificación o regresión, esta técnica de ensambles combina los resultados de varios modelos de árboles de decisión para obtener un resultado definitivo, se pueden utilizar decenas o centenas de árboles para conformar el RandomForest, aunque para detectar el número óptimo de árboles se suele utilizar la validación cruzada.

Para la construcción del RandomForest se puede utilizar una librería de Scikit-learn RandomForestRegressor, ya que se aplicará a un problema de regresión.

Entre los parámetros del modelo de RandomForests requiere indicar el número de estimadores el cual es un parámetro que indica la cantidad de árboles que tendrá el modelo, para optimizar este parámetro, se hace uso de la técnica “Cross Validation”, que permite dividir los índices del Dataset en entrenamiento y prueba, lo que permite evitar obtener los parámetros con los mismos datos de entrenamiento protegiendo al modelo del Overfitting, con el uso de la estrategia “k-fold” para dividir el Dataset en varios pliegues de forma aleatoria para obtener el mejor estimador utilizando distintas potencias del 2 hasta un máximo de número de estimador de 1024. (Pedregosa, 2011)

- **n_estimators:**
  Indica el número de árboles que tendrá el modelo.

- **criterion:**
  Indica la función con la que se medirá la calidad de cada división.

- **max_depth:**
  Indica la profundidad máxima de los árboles, por defecto este parámetro viene en estado “None”, lo que indica que los nodos del árbol seguirán creciendo hasta que todos los nodos contengan menos del mínimo de ejemplos para realizar una división.

- **random_state:**
  Controla la aleatoriedad de las muestras utilizadas al construir árboles y el muestreo de las características, uno de los valores más populares es el 0 que indica el generador de una semilla para que los datos sean reproducibles.

- ***DecisionTree***

El árbol de decisión es un modelo de Machine Learning donde básicamente se genera una serie de preguntas que se pueden considerar nodos los cuales crean caminos o bifurcaciones con más preguntas, otorgando así su nombre de árbol de decisión, puede ser utilizado para tareas de clasificación o regresión, la profundidad del árbol será el número de nodos que posee.

El árbol de decisión es un modelo muy sencillo de predicción, para definir la mejor profundidad del árbol se hace uso de la técnica Cross Validation para iterar entre varias profundidades del árbol entre el rango de valores 2 y 50, y obtener el menor MAE, el árbol de decisión, aunque es sencillo de entender, se torna complicado cuando este posee mucha profundidad, por esta razón no se utilizan las potencias del 2 para definir la máxima profundidad en esta ocasión. (Pedregosa, 2011):

- **max_depth:**
  Máxima profundidad del árbol.

- **criterion:**
  Indica la función con la que se medirá la calidad de cada división.

- ***AdaBoost***

La técnica “AdaBoost” es un algoritmo de Boosting adaptativo que, mediante un entrenamiento de clasificadores débiles, asigna mayor importancia a los datos mal clasificados en la anterior iteración lo que le permite ir adaptándose y obteniendo un nuevo clasificador más apto cada vez, este tipo de conjunto de modelos suele dar mejores o por lo menos los mismos resultados que un solo regresor, se puede utilizar la técnica “AdaBoost” para impulsar un árbol de decisión (Pedregosa, 2011).

- **n_estimators:**
  Indica el número máximo de estimadores donde finalizara el refuerzo.

- **learning_rate:**
  Indica la tasa de peso que se aplica a cada regresor en cada iteración, cuanto más grande es el número más alto es la ratio de aprendizaje, por defecto este valor es de 1.0.

- **loss:**
  Es la función que se utiliza para actualizar los pesos de cada iteración, por defecto la función que se utiliza es “linear”.

- **random_state:**

Controla la semilla dada a cada base_estimator en cada iteración, también controla el arranque de los pesos utilizados para entrenar cada estimador en cada iteración.

- ***GradientBoosting***

Similar a la técnica AdaBoost, el GradientBoosting es método que utiliza como base los árboles de decisión, ya que estos modelos compuestos de clasificadores débiles suelen obtener resultados favorables de forma secuencial, el algoritmo permite que cada modelo aprenda de los errores del modelo anterior, el tercer modelo intentara aprender del error generado por los dos modelos anteriores y así sucesivamente, este proceso reduce los residuos o errores en cada iteración por lo que puede darse el problema de que el modelo sobreentrene o aprenda los datos de memoria (Overfitting), por lo tanto se debe ajustar un valor de regulación que limita la influencia de cada modelo del conjunto de modelos, este regulador se conoce como "learning rate", el cual es un valor que se indica como hiperparámetro y puede ir en un rango de 0 a "inf" aunque se recomienda seleccionar valores entre 0.001 y 0.01 (Amat, 2020).

- **loss:**
  Indica la función de perdida para ser optimizada.
- **learning_rate:**
  Indica la reducción de tasa de aprendizaje de cada árbol en cada iteración, existe una relación negativa entre este parámetro y el parámetro n_estimators, su valor por defecto es de 0.1.
- **n_estimators:**
  Indica el número de etapas de refuerzo a realizar.
- **random_state:**
  Controla la semilla dada a cada estimador en cada iteración, además de controlar la división aleatoria de los datos de entrenamiento.
- **criterion:**
  Indica la función con la que se medirá la calidad de cada división.

5. Desarrollo del proyecto

5.1. Tipos de metodología

Existen diversas metodologías relacionadas con la minería de datos, algunas de ellas son extensiones de otras metodologías, que describen procesos o tareas que el analista o científico de datos debe recorrer para construir un proyecto de ciencia de datos, o de Big Data. Las metodologías de minería de datos suelen comenzar con el entendimiento del negocio y finalizar con el despliegue del desarrollo que se halla implementado, en caso de que el proceso no haya sido meramente analítico, las metodologías suelen tener tareas similares con algunas variaciones, sin embargo, todas conservan la idea fundamental de la minería de datos la cual es la de descubrir significancia, patrones o tendencias de un conjunto de datos para darles valor. (Hernández G & Dueñas R., 2009)

A continuación, se describen tres metodologías utilizadas en la minería de datos:

5.1.1. CRISP-DM

CRISP-DM (Cross Industry Standard Process For Data Mining) es una metodología que se originó en los años 90 con el fin de normalizar el proceso KDD (Knowledge Discovery in Databases), la metodología CRISP-DM especifica una lista de tareas que se deben realizar en cada fase para conseguir realizar un objetivo, esta metodología es popular ya que tiene en cuenta aplicaciones a entornos de negocio, el ciclo de vida es de 6 fases, donde cada fase cumple con una o varias tareas definidas y dependiendo del resultado se determina con que fase se debe continuar. (SNG)

La metodología CRISP-DM posee un ciclo de vida con seis fases, las cuales interactúan entre ellas para optimizar el desarrollo del proyecto, las fases de la metodología CRIPS-DM son las siguientes:

- ***Fase I. Business Understanding.***
  La primera fase se indaga en conocer los objetivos y la organización del proyecto de acuerdo con las necesidades del negocio, en esta fase se suele involucrar personal de negocio para obtener una visión general del problema.

- ***Fase II. Data Understanding.***
  La fase dos se encarga de la recolección de datos que permitan la ejecución de proyecto y sean la base principal del mismo.

- ***Fase III. Data Preparation.***
  Una vez recolectados los datos que servirán como base del proyecto, se debe realizar una limpieza para eliminar datos perdidos, datos fuera de rango (Outliers), y datos repetidos o inconsistentes, también se realiza el proceso de transformación de datos si es necesario.

- ***Fase IV. Modeling***
  En la fase del modelado se seleccionan las técnicas adecuadas para realizar la minería de datos, las técnicas de modelado se implementarán dadas las condiciones y características de los datos.

- ***Fase V. Evaluation***
  La fase de evaluación sirve para verificar los resultados del modelaje y las etapas anteriores, si los resultados son satisfactorios se avanza a la última fase, de lo contrario se debe regresar a las fases anteriores y realizar correcciones.

- ***Fase VI. Deployment***
  La fase final consiste en llevar el proyecto a producción dentro de organización, donde se realiza la entrega de resultados y documentación sobre la aplicación.

5.1.2. SEMMA

La metodología SEMMA (Sample, Explore, Modify, Model, Assess), se define como un proceso de selección, exploración y modelado utilizando grandes cantidades de datos que permitan el descubrimiento de patrones que sirvan como herramienta de apoyo para el negocio, se define como un conjunto de enfocadas en el desarrollo de minería con cinco fases de desarrollo. (Hernández G & Dueñas R., 2009)

- ***Muestreo***
  El primer paso es el muestreo o extracción de una muestra de datos que posea características representantes de una población, reduciendo costes y tiempo para la organización.

- ***Exploración***
  La exploración de datos permite identificar y eliminar datos perdidos, fuera de rango, o con anomalías o deficiencias.

- ***Modificación***
  En esta fase se realiza el proceso de transformación de datos dependiendo de las características que se seleccionan para el proceso de minado de datos, en la siguiente fase.

- ***Modelado***
  En este punto de la metodología utilizan herramientas de software para emplear técnicas de minería de datos que permitan descubrir o patrones entre los datos seleccionados para realizar predicciones, las técnicas pueden incluir métodos estadísticos, redes neuronales, arboles de decisión y lógica difusa entre otros.

- ***Evaluación***
  El paso final consiste en validar los resultados obtenidos a partir de los modelos utilizados en la fase anterior. Cuando se habla de predicciones, se debe dividir

el conjunto de datos de tal forma que se pueda realizar el modelo con la parte más grande y dejar la sección más pequeña para realizar la validación de la efectividad del modelo, en caso de que los resultados sean óptimos se lleva a producción el modelo, de lo contrario se debe regresar al paso anterior para cambiar el modelo seleccionado.

5.1.3. ASUM-DM

La metodología ASUM-DM fue creada a partir de la metodología CRISP-DM, enfatiza nuevas técnicas de ciencia de datos. ASUM-DM se concentra en 5 grupos de fases globales (1. Analizar,2. Diseñar, 3. Configurar y Construir, 4. Desplegar, Operar, Optimizar). Dentro de cada grupo se despliegan las 10 fases que componen la metodología las cuales se describen a continuación (ANDES, 2017).

- ***Comprensión del negocio***
  El primer paso establece las bases para abordar el proyecto estableciendo pautas que ayuden a definir el entendimiento de negocio.

- ***Enfoque analítico***
  Una vez se ha traducido el problema de negocio a un problema técnico, se pueden definir el enfoque analítico para resolver el problema, seleccionado las herramientas estadísticas o de aprendizaje automático.

- ***Requisitos de los datos***
  Con la selección del enfoque analítico se determinan los requisitos que deben tener los datos para la construcción del proyecto.

- ***Entendimiento de los datos***
  Utilizando técnicas de visualización o estadística permiten la comprensión de la base de datos para evaluar la calidad de los mismos y posiblemente realizar hallazgos de valor para el proyecto.

- ***Preparación de los datos***
  En esta etapa del proyecto se realizan las tareas referentes limpieza de datos y transformación de datos si es necesario para crear robustes, en el conjunto de datos al que se le aplicaran los modelos en la siguiente etapa.

- ***Modelamiento***
  Con el conjunto de datos limpio y transformado se entra en la etapa de modelado donde se hace uso de técnicas de Machine Learning para la realización de predicciones de características específicas, según las necesidades del negocio que previamente hayan sido establecidas.

- ***Evaluación***
  Se realiza la evaluación de los modelos creados en la etapa anterior, y se examina la calidad de los mismos utilizando diversas medidas de diagnóstico que permitan asegurarse de que los modelos abordan correctamente la problemática del negocio.

- ***Despliegue***
  Si la prueba de evaluación fue exitosa, se procede a desplegar el aplicativo en el entorno de producción, de lo contrario se debe regresar a alguna de las etapas anteriores para realizar ajustes pertinentes.

- ***Retroalimentación***
  Se recolectan datos sobre el modelo puesto en producción, para obtener retroalimentación sobre el rendimiento del modelo dicho ambiente.

5.2. Selección de metodología

Como se observó en la sección anterior, existen diversas metodologías que se pueden aplicar a los proyectos relacionados con el análisis de datos. Las tres metodologías comparten similitudes en algunos aspectos, variando en la cantidad de pasos o fases que cada metodología posee. Básicamente todas las metodologías poseen una sección de entendimiento del negocio, una sección de recolección, una sección de limpieza/transformación de datos y finalmente la sección de modelado y evaluación. En casos donde la necesidad del negocio es teórica como en el caso de este proyecto, no se requeriría la sección de despliegue, también se debe aclarar que la metodología ASUM-DM fue creada a partir de la metodología CRISP-DM, acoplando la sección de retroalimentación posterior al despliegue, pero como se mencionó antes esta sección no aplicaría en este proyecto. Por último, la metodología SEMMA, resulta muy similar a la metodología CRISP-DM con la diferencia de que, en vez de tomar toda la cantidad de datos disponibles, se toma una muestra representativa en la primera fase, esta metodología incluye el entendimiento del negocio en la fase de recolección de datos. Para este proyecto se planea utilizar la mayor cantidad de datos disponibles dado que las ligas femeninas no poseen la misma cantidad de registros sobre encuentros deportivos como las ligas masculinas de las que si hay mucha información, y como una de las metas de este proyecto es realizar un estudio estadístico a las ligas femeninas, se debe contar con la mayor cantidad de datos posibles, razón por la cual se decidió utilizar la metodología CRISP-DM para la realización de este proyecto.

5.3. Aplicación de metodología CRISP-DM

- **Fase I. Business Understanding.**

En la primera fase se comprenderán los objetivos establecidos para este proyecto, con el fin de generar un plan de desarrollo incluyendo la definición de fuentes de donde se realizará la obtención de los datos con los cuales se realizará el estudio, para el caso

de este proyecto, se indagara sobre el efecto que existe en los equipos deportivos al cambio de entrenador, y para ello se definirán las fuentes de datos donde sea posible encontrar fechas de entrada y salida de entrenadores de clubes deportivos de futbol en diversas ligas masculinas y femeninas.

- **Fase II. Data Understanding.**

La fase dos está destinada a la recolección de los datos iniciales que se utilizaran para conseguir los objetivos propuestos, en este proyecto hay variedad de fuentes para tomar datos, sin embargo, se sabe que el problema real será el de unificar la información en un solo Dataset que contenga fechas de inicio y fin en que un entrenador está contratado, jornadas enteras con información del estado de los equipos deportivos en ese momento y marcadores exactos de los encuentros, una vez se recolecten estos datos, se empezara a descubrir información valiosa que ayudara a comprender mejor el efecto de cambio de entrenador.

- **Fase III. Data Preparation**.

En esta fase se realizan todas las actividades correspondientes a la preparación de los Datasets obtenidos, se realizara la limpieza de los mismos, los valores faltantes de fechas serán completados con otros valores y en el caso de los datos con marcadores faltantes, se deben remover ya que los datos “NaN” no contribuyen a la construcción de modelos, no vale la pena rellenar estos espacios con la media o mediana de los datos, ya que cada enfrentamiento es un caso independiente, posteriormente se transformaran los Datasets para poder fusionarlos y posiblemente preparar un único Dataset que contenga toda la información, así será posible manipularlo con mayor facilidad, una vez preparados los datos, se puede utilizar el Dataset limpio para estudiar las medias de las variables y visualizar el efecto de cambio de entrenador utilizando la regresión lineal como herramienta.

- **Fase IV. Modeling.**

En esta fase se realizaran dos procesos, el primero será el estudio estadístico utilizando la regresión lineal en el Dataset limpio con el fin de evidenciar el posible efecto positivo en el rendimiento de un equipo después de que este realiza un cambio de entrenador, en las diferentes ligas masculinas y femeninas, el segundo proceso será el de construir un modelo de Machine learning sobre los datos normalizados buscando predecir el promedio de puntos a corto plazo que un equipo deportivo puede obtener en los encuentros posteriores al cambio de director deportivo.

- **Fase V. Evaluation.**

En esta sección se compararán los resultados obtenidos en el estudio estadístico de los datos de las ligas masculinas y femeninas de primera división también se compararán los modelos creados de Machine Learning para la predicción del promedio de puntos obtenidos a corto plazo luego del cambio de entrenador, como métricas de error se utilizará el MAE (índice de error absoluto medio) y el MSE (error cuadrático medio).

- **Fase VI. Deployment.**

Para este proyecto, la fase del despliegue no se realizará ya que es un proyecto teórico.

- **Conclusión.**

El paso final es la conclusión con base en los resultados obtenidos en la fase IV y la evaluación realizada en la fase V.

5.4. Planteamiento del problema

El cambio de entrenador en un equipo deportivo es un tema que se ha estudiado por muchos años, llevando a los investigadores a crear varias teorías de las posibles causas del cambio de entrenador y del impacto que se genera en los equipos a corto y largo plazo luego del cambio, estas investigaciones se han dado estudiando ligas masculinas de primera división y en algunas ocasiones de segunda división. No se han encontrado investigaciones en el estado del arte que utilicen datos de ligas femeninas, esto puede deberse a la bastedad de datos disponibles que ofrecen las ligas masculinas.

Hoy en día, las ligas femeninas han comenzado a tener más popularidad que en antaño, lo que permite contar con una mayor cantidad y calidad de datos. Ya se ha investigado el tema del cambio de entrenador en ligas masculinas anteriormente, donde varios de los autores vistos en el estado del arte apoyan la primera teoría del cambio de entrenador (The common-sense one-way casuality theory), pero no está claro si para las ligas femeninas ocurre el mismo efecto, también se debe destacar que los estudios realizados fueron hechos en ligas separadas, como la Premier League o la Bundesliga. Se observó que los equipos a largo plazo mejoraran, empeoraran o se mantendrán igual que antes del cambio de entrenador, pero esta consecuencia no tiene que ver con la llegada del entrenador, sino con el estilo de juego que el entrenador proponga y el acoplamiento que el equipo tenga al estilo de juego propuesto, lo cual conllevara mucho entrenamiento y adaptación, pero las directivas y en general la fanaticada preferirán cambios inmediatos si su club deportivo está obteniendo malos resultados, razón por la cual se crearon las tres teorías del cambio de entrenador. Para determinar si existen efectos similares en las ligas femeninas se debe realizar el estudio de los resultados de los equipos después del cambio de entrenador a corto plazo, y compararlo con los resultados obtenidos por las ligas masculinas.

5.5. Aplicación de la metodología CRIPS-DM

1. Fase I. Business Understanding

Como se ha visto en el estado del arte, existen investigaciones dedicadas al entendimiento sobre el posible efecto que existe cuando un equipo decide cambiar a su entrenador, algunos autores estudian únicamente las causas por las cuales se destituye un entrenador y otros observan el comportamiento que tiene el equipo luego de este suceso, gracias a estos análisis fue posible la creación de las 3 teorías sobre el cambio de entrenador, en este proyecto se busca realizar un estudio similar analizando algunas de las mejores ligas masculinas y femeninas de futbol del mundo según (As, s.f.)

Para llevar a cabo el estudio, se escogieron las siguientes ligas de fútbol profesional masculino:

- Betplay Dimayor (Colombia, Primera división)
- Brasileirao (Brasil, Primera división)
- Bundesliga (Alemania, Primera división masculina)
- Eredivisie (Países Bajos, Primera división masculina)
- LaLiga (España, Primera división masculina)
- Ligue 1 (Francia, Primera división masculina)
- Premier League (Inglaterra, Primera división masculina)
- Primeira liga (Portugal, Primera división masculina)
- Serie A (Italia, Primera división masculina)
- Primera división de Argentina (Argentina, Primera división)

Las anteriores ligas están ordenadas en orden alfabético, dentro de estas ligas, 7 pertenecen a Europa y 3 a Latinoamérica, lo cual es beneficioso para este proyecto ya que es posible realizar una comparación entre continentes, aprovechando que ambos poseen excelentes ligas de estudio.

Las ligas femeninas europeas de futbol no han tenido tanta fama como las ligas masculinas razón por la cual no se han encontrado estudios sobre la temática de cambio de entrenador utilizando ligas femeninas. Se seleccionaron algunas de las mejores ligas femeninas para realizar un análisis y poder comparar los resultados con los resultados de las divisiones masculinas europeas y latinoamericanas según (90min, s.f.)

- Bundesliga (Alemania, Primera división femenina)
- Damallsvenskan (Suiza, Primera división femenina)
- Eredivisie (Países Bajos, Primera división femenina)
- LaLiga (España, Primera división femenina)
- Ligue 1 (Francia, Primera división femenina)
- Liga MX femenil (México, Primera división femenina)
- National Women´s Soccer League (Estados Unidos, Primera división femenina)
- Premier League (Inglaterra, Primera división femenina)
- Toppserien (Noruega, Primera división femenina)

Con la recolección de datos sobre estas diferentes ligas tanto masculinas como femeninas se busca realizar un análisis estadístico, que abarque los equipos pertenecientes a todas las anteriores ligas y permita encontrar comparaciones y similitudes; Para alcanzar esta meta se debieron encontrar fuentes de datos con los registros de los resultados de muchos años, y con estos resultados se debieron encontrar también los directores técnicos que se encontraban o encuentran vigentes en dichos equipos, para poder reconocer cuando estos son reemplazados y estudiar el fenómeno de cambio de entrenador: Se requiere de fechas en las que los entrenadores toman y dejan sus puestos, es importante tomar las fechas exactas del cambio ya que se podría tomar el entrenador por temporada, sin embargo, existe el problema de que un club deportivo puede cambiar de entrenador muchas veces en una sola temporada, por lo que sería difícil elegir un entrenador, además el ideal es poder examinar a los entrenadores una vez llegan al equipo, si se tomase por temporada se estaría tomando una perspectiva herrada de los reales resultados, por esta razón es muy importante tomar las fechas de entrada y salida en que un entrenador dirige un equipo.

Para la recolección de datos se tendrá encuentra toda la cantidad de partidos disponible de cada liga en la página “livefutbol.com” (livefutbol, s.f.). Las temporadas para cada liga son las siguientes:

- Betplay Dimayor: 2007-2022
- Brasileirao:1995-2022
- Bundesliga: (1963/1964) - (2022/2023)
- Eredivisie: (1956/1957) - (2022/2023)
- LaLiga: (1928/1929) - (2022/2023)
- Ligue 1: (1932/1933) - (2022/2023)
- Premier League: (1888/1889) - (2022/2023)
- Primeira Liga: (1955/1956) – (2022/2023)
- Serie A: (1929/1930) – (2022/2023)
- Primera división de Argentina: (1985/1986) – (2022/2023)

Las temporadas disponibles para las ligas femeninas son las siguientes:

- Bundesliga: (1997/1998) - (2022/2023)
- Damallsvenskan (2012/2013) - (2022/2023)
- Eredivisie: (2007/2008) - (2022/2023)
- LaLiga: (2013/2014) - (2022/2023)
- Ligue 1: (2013/2014) - (2022/2023)
- Liga MX femenil (2018/2019) – (2022/2023)
- National Women´s Soccer League (2013/2014) – (2022/2023)
- Premier League: (2018/2019) – (2022/2023)
- Toppserien (2009/2010) – (2022/2023)

La mayoría de ligas femeninas seleccionadas son de Europa exceptuando las ligas de México y estados unidos, dado que suelen haber pocos datos sobre las ligas femeninas, se incluyeron estas dos ligas adicionales a las ligas europeas femeninas para aumentar la cantidad de datos disponibles para realizar el análisis.

2. Fase II. Data Understanding

- ***WebScraping para construcción de Datasets***

Para la realización de este proyecto fue necesario obtener una gran cantidad de datos relacionados con las ligas masculinas y femeninas de futbol, por tanto, se construyó un código en Python para obtener la información de una página de deportes página "livefutbol.com" (livefutbol, s.f.). Es indispensable la obtención de algunos datos importantes como la fecha en la que se jugó un partido, la liga en que se jugó, el número de goles realizado por cada equipo y el entrenador vigente en cada equipo durante ese partido. Lograr obtener o descargar un Dataset con estas características en particular es difícil ya que las páginas oficiales de los clubs pueden o no tener esta información, y en caso de que tengan los partidos y resultados, no es seguro encontrar los entrenadores vigentes en dichos partidos exactamente, ya que se encuentran archivados por temporadas, esto puede ser un problema ya que, en una temporada, un entrenador puede ser reemplazado sin haber terminado la temporada, para ello la estrategia que se siguió fue la de escoger de una liga en particular, todos los partidos de cada temporada y de cada equipo, obtener el resultado de ese partido y la fecha en la que se jugó.

Posteriormente se realizó un proceso para obtener todos los entrenadores que ha tenido un club de futbol, junto con la fecha de ingreso del entrenador a ese club y la fecha de salida, este proceso fue largo, ya que se realizó de igual forma para todos los equipos que se hallan jugado en la primer división de futbol de las ligas seleccionadas, una vez realizado este proceso y obtenido estos Datasets por separado, se fusionaron en un solo conjunto de datos, con el fin de que en cada partido se pueda visualizar el entrenador que estaba vigente en ese partido, además de las columnas de las fechas y los resultados, de esta manera si un entrenador renuncio o fue despedido en medio de la temporada, se podrá visualizar claramente que entrenador lo está reemplazando.

Primero se debe visualizar la página de la que se van a obtener los datos y la forma en la que vienen acomodados, se debe seleccionar el país del que se quiere obtener toda la información.

Conociendo la estructura de la página, fue posible realizar un programa en Python con la ayuda de la librería "Beautiful Soup" (Librería enfocada en el Webscraping) que recorre cada pestaña de cada temporada y luego para cada temporada recorre cada jornada, así se obtuvo una tabla con los valores de cada partido en un rango de años, luego estos valores se almacenaron en un DataFrame que se guardó en un archivo de tipo CSV.

Para cada una de las ligas seleccionadas en la fase I, se generó un Dataset individual, es decir que en total hay 10 Datasets de la liga masculina primera división y 9 Datasets de la liga femenina primera división, las ligas masculinas son muy antiguas lo cual permitió tener un rango de estudio más elevado, por ejemplo, la Premier League posee

registros desde el año 1889, o Laliga donde existen registros desde 1928, desafortunadamente para el caso de las ligas femeninas hay mucha menos información de la que se puede disponer, apenas hay registros del año 2013 para Laliga femenina en la página “livefutbol.com” (livefutbol, s.f.), y del 2018 para la Premier League femenina, razón por la cual se seleccionaron datos de varias ligas femeninas de primera división incluso si no se consideran de las mejores ligas profesionales, como es el caso de la liga de México femenil.

El resultado de estos Datasets tiene la siguiente estructura de variables:

- fecha: La fecha donde se jugó el partido.
- jornada: El número de la jornada en que se jugó el partido.
- local: El equipo que jugó con condición de local.
- visitante: El equipo que jugó en condición de visitante.
- goles_local: Cantidad de puntos o goles anotados por el equipo local.
- goles_visitante: Cantidad de puntos o goles anotados por el equipo visitante.

Cada dataset posee las siguientes dimensiones:

Ligas masculinas:

- betplay_dimayor_results: 5543 filas por 6 columnas.
- brasileirao_results: 10743 filas por 6 columnas.
- bundesliga_results: 18302 filas por 6 columnas.
- eredivisie_paises_bajos_results: 19932 filas por 6 columnas.
- laliga_results: 26576 filas por 6 columnas.
- ligue_1_results: 29648 filas por 6 columnas.
- premier_league_results: 50190 filas por 6 columnas.
- primeira_liga_portugal_results: 17462 filas por 6 columnas.
- primera_division_argentina_results: 13994 filas por 6 columnas.
- serie_a_italia_results: 28104 filas por 6 columnas.

Ligas femeninas:

- bundesliga_femenina_results: 3432 filas por 6 columnas.
- Damallsvenskan_suecia_femenina_results: 1898 filas por 6 columnas.
- eredivisie_paises_bajos_femenina_results: 1007 filas por 6 columnas.
- laliga_femenina_results: 2226 filas por 6 columnas.
- ligue_1_femenina_results: 1584 filas por 6 columnas.
- mex_primera_division_femenina_results: 1530 dilas por 6 columnas.
- National_womens_soccer_league_femenina_results: 1069 filas por 6 columnas.
- premier_league_femenina_results: 638 filas por 6 columnas.
- toppserien_noruega_femenina_results: 1734 filas por 6 columnas.

- ***WebScraping para Datasets de los entrenadores***

Como se pudo observar en la sección anterior, no existe por el momento una columna con el indicador del entrenador vigente en cada partido disputado, esta información se debe añadir ya que es una de las columnas más importantes, por lo tanto, se realizó un programa en Python para recolectar el entrenador por equipo y por fecha de entrada y salida del equipo, para luego incluir esta información a los Dataset creados anteriormente.

Para obtener esta información se utilizó una técnica de WebScraping similar a la que se utilizó en los Datasets de los resultados, donde se utilizó el valor único de cada club que ha participado en una liga en particular para buscar todos los entrenadores de dicho club y así obtener su nombre y la fecha en que estuvo vigente.

Se creo un Dataset para cada liga, con la información de los entrenadores, cada Dataset posee las siguientes variables:

- club: Club deportivo de futbol que dirigió el entrenador.
- fecha_activo_inicio: Fecha en la que inicio el entrenador en el club.
- fecha_activo_fin: Fecha en la que finalizo sus servicios el entrenador en ese club.
- entrenador: Nombre del entrenador.
- país: país de origen del entrenador.
- fecha_nacimiento: Fecha de nacimiento del entrenador.

Cada Dataset posee las siguientes dimensiones:

Entrenadores ligas masculinas:

- entrenadores_betplay_dimayor: 592 filas por 6 columnas.
- entrenadores_brasileirao: 1580 filas por 6 columnas.
- entrenadores_bundesliga: 3224 filas por 6 columnas.
- entrenadores_eredivisie: 1997 filas por 6 columnas.
- entrenadores_laliga: 2867 filas por 6 columnas.
- entrenadores_ligue_1: 1929 filas por 6 columnas.
- entrenadores_premier_league: 2471 filas por 6 columnas.
- entrenadores_primeira_liga: 1347 filas por 6 columnas.
- entrenadores_primera_division_argentina: 1529 filas por 6 columnas.
- entrenadores_serie_a: 5769 filas por 6 columnas.

Entrenadores ligas femeninas:

- entrenadores_bundesliga_femenina: 326 filas por 6 comulnas.
- entrenadores_damallsvenskan_suecia_femenina: 77 filas por 6 columnas.
- entrenadores_eredivisie_femenina: 66 filas por 6 columnas.
- entrenadores_laliga_femenina: 46 filas por 6 columnas.
- entrenadores_ligue_1_femenina: 64 filas por 6 columnas.
- entrenadores_mex_primera_division_femenina: 56 filas por 6 columnas.

- entrenadores_national_womens_soccer_league_femenina: 74 filas por 6 columnas.
- entrenadores_premier_league_femenina: 42 filas por 6 columnas.
- entrenadores_toppserien_noruega_femenina: 17 filas por 6 columnas.

- ***WebScraping para datos obtenidos de La FIFA***

Para complementar los datos obtenidos anteriormente, se realizó una búsqueda en la página "*https://www.fifaindex.com*" (fifaindex, s.f.) donde se encuentran características acerca de la defensa y el ataque de diversos equipos de varias ligas, también se pueden encontrar datos sobre características de jugadores de dichos equipos, esta información proviene de los videojuegos de la FIFA desde el año 2005 hasta el presente año.

Nuevamente se utilizó la librería "Beaufitul Soup" para realizar WebScraping en la página de La FIFA y obtener un Dataset que contenga las características de cada equipo en cada año desde el 2005.

El Dataset obtenido tiene la siguiente estructura de variables:

- equipo: Nombre del equipo.
- liga: Liga a la que pertenece el equipo.
- ataque: Valoración de puntos de habilidad en ataque del equipo.
- medio: Valoración de puntos de habilidad en el medio campo.
- defensa: Valoración de puntos de habilidad en defensa del equipo.
- promedio: Promedio de puntos entre el ataque, medio y defensa del equipo.
- año: Año del FIFA correspondiente.

Se genera un Dataset llamado "fifa_caracteristicas_equipos" contiene 10350 filas por 7 columnas.

3. Fase III. Data Preparation.

- ***Acumulación de Datasets por Continente y género.***

En la fase anterior se realizó Webscraping para obtener datos de las mejores ligas profesionales de futbol del mundo, incluyendo algunas ligas latinoamericanas y también algunas ligas europeas femeninas, con el fin de visualizar el efecto de cambio de entrenador en diferentes regiones y también verificar si el efecto se reproduce en las ligas femeninas.

En esta fase se realizaron las tareas de transformar y limpiar los Datasets para luego unirlos en un conjunto. Se debe recordar que hasta el momento se cuenta con 38 Datasets, 19 Datasets con datos relacionados a los partidos jugados de las ligas

mencionadas, incluyendo 9 ligas europeas femeninas, y 19 Datasets con la lista de los entrenadores para cada equipo de cada liga de las ligas seleccionadas; Fue necesario unificar estos archivos para así poder prepararlos para la siguiente fase.

En primer lugar, se unieron los Dataframes de las ligas europeas masculinas en un solo Dataset, aunque antes de realizar este proceso se incluyó una nueva columna a todos los Datasets existentes con la liga a la que pertenecen, es decir que ahora todos los Datasets poseen 7 columnas.

Se unieron todos los Datasets referentes a las ligas masculinas de Europa en un solo Dataset, se repite el proceso para las ligas europeas femeninas, luego se repite el proceso para las ligas latinoamericanas, lo que resulta en la creación de tres nuevos Datasets.

Estructura de los Datasets con acumulación de encuentros:

- fecha: La fecha donde se jugó el partido.
- jornada: El número de la jornada en que se jugó el partido.
- local: El equipo que jugó con condición de local.
- visitante: El equipo que jugó en condición de visitante.
- goles_local: Cantidad de puntos o goles anotados por el equipo local.
- goles_visitante: Cantidad de puntos o goles anotados por el equipo visitante.
- liga: Liga a la que pertenece el encuentro.

Dimensiones de los Datasets con acumulación de encuentros:

- europa_results: 190214 filas por 7 columnas.
- europa_results_femenina: 15490 filas por 7 columnas.
- latinoamerica_results: 30280 filas por 7 columnas.

Se realizó el mismo proceso de acumulación para los Datasets de los entrenadores, se unieron todos los Datasets de los entrenadores de las ligas masculinas de Europa, se repitió el proceso para los entrenadores de las ligas femeninas de Europa. Finalmente se unieron todos los Datasets de los entrenadores de las ligas latinoamericanas, lo que resulta en la creación de 3 Datasets nuevos con la acumulación de los entrenadores.

Estructura de los Datasets con acumulación de entrenadores:

- club: Club deportivo de futbol que dirigió el entrenador.
- fecha_activo_inicio: Fecha en la que inicio el entrenador en el club.
- fecha_activo_fin: Fecha en la que finalizo sus servicios el entrenador en ese club.
- entrenador: Nombre del entrenador.
- país: país de origen del entrenador.
- fecha_nacimiento: Fecha de nacimiento del entrenador.
- liga: Liga a la que pertenece el entrenador en ese momento.

Dimensiones de los Datasets con acumulación de entrenadores:

- europa_managers: 19605 filas por 7 columnas.

- europa_managers_femenina: 770 filas por 7 columnas.
- latinoamerica_managers: 3699 filas por 7 columnas.

- ***Transformación de Datasets de resultados para unificar encuentros con Datasets de entrenadores.***

Se han acumulado todos los Datasets de las ligas de Europa en uno solo, continuando de igual manera con los Datasets de Latinoamerica, Europa femenina, entrenadores de Europa, entrenadores de Europa femenina y Entrenadores de Latinoamérica, en esta sección se realizó un proceso de transformación en el Dataset de resultados para integrarlo con el Dataset de entrenadores con el fin de lograr poner en cada resultado, los entrenadores del equipo local y el equipo visitante que en ese momento dirigían el equipo.

Se completaron las fechas faltantes que poseen los resultados, como se mencionó anteriormente, es posible completar estos datos con el dato anterior, ya que los partidos de una jornada en particular suelen jugarse el mismo día o en días cercanos, un día o dos adelante o un día o dos atrás, por ende, no es significativa la diferencia de fechas tan solo por uno o dos días. El proceso se realizó creando una función en Python que recorra el Dataset de los resultados y cada vez que encuentre un valor “NaN” en la columna de “fecha”, lo complete con el último dato de fecha que haya guardado, el mismo proceso se aplicó para el Dataset referente a los entrenadores.

Como ya se encuentran completas las filas de la columna fechas, se realizó la transformación del Dataset, para ello se debió aplicar un procedimiento comparando la fecha del encuentro con el rango de fechas en que un entrenador dirigió un club deportivo, es un proceso muy largo ya que cada fila de los Datasets de entrenadores debe recorrer todo el Dataset de lo resultados y comparara si el equipo local es igual al equipo del entrenador, de ser así se comparara la fecha del encuentro con el rango de fechas de entrada y salida del entrenador, los nombres de los entrenadores se irán acomodando en una nueva columna junto con el país de origen del entrenador, al terminar el bucle, se incluirán estas listas como columnas al Dataset de resultados, y enseguida se repitió el mismo proceso pero esta vez para los entrenadores de los equipos visitantes, es decir que 4 nuevas columnas serán agregadas a los Datasets de resultados, este proceso se realizara de la misma forma para los Datasets “europa_results_femenina” y “latinoamerica_results” con los Datasets de entrenadores llamados “europa_managers_femenina” y “latinoamerica_managers” respectivamente.

La estructura de los Datasets de resultados unificados con entrenadores es la siguiente:

- fecha: La fecha donde se jugó el partido.
- jornada: El número de la jornada en que se jugó el partido.
- local: El equipo que jugó con condición de local.
- visitante: El equipo que jugó en condición de visitante.
- goles_local: Cantidad de puntos o goles anotados por el equipo local.

- goles_visitante: Cantidad de puntos o goles anotados por el equipo visitante.
- liga: Liga a la que pertenece el encuentro.
- entrenador_local: Nombre del entrenador local.
- pais_entrenador_local: País donde nació el entrenador local.
- entrenador_visitante: Nombre del entrenador visitante.
- pais_entrenador_visitante: País donde nació el entrenador visitante.

Dimensiones de los nuevos Datasets unificados.

- europa_resultados: 190214 filas por 11 columnas.
- europa_femenina_resultados: 15490 filas por 6 columnas.
- latinoamerica_resultados: 30280 filas por 11 columnas.

- ***Limpieza de Datasets unificados***

Se ha unificado toda la información de las ligas y entrenadores en los anteriores 3 Datasets, separados por continente y género completando las fechas faltantes en todos los Datasets antes de unificarlos, sin embargo, siguen existiendo algunos datos faltantes como goles en partidos que aún no se han disputado al momento de realizar este proyecto, por lo tanto, se debe remover de los Dataset las filas con estos datos “NaN”.

Existen algunos datos faltantes en las columnas de los goles y en la columna de los nombres de entrenadores con sus respectivos países, los goles pertenecen a fechas que aún no se han disputado al momento de realizar este proyecto, por ejemplo, si observamos los datos para la jornada número 34 de la Bundesliga masculina de la temporada 2022/2023, observaremos que este enfrentamiento aún no sucede, por lo tanto, se removieron las filas que no contengan encuentros disputados ya que no ofrecerán ningún dato valioso.

Los datos faltantes pertenecientes a los entrenadores local y visitante junto con los países a los que pertenecen no están disponibles en el Dataset debido a que al momento de realizar el Webscraping no se encontraban en la página, por ejemplo, el club deportivo “Toulouse” Ligue_1 división femenina, no posee registro de entrenadores.

Tener presente la información sobre cual entrenador estaba dirigiendo a un equipo en el momento de un encuentro es muy importante para este proyecto. Es necesario remover las filas que contengan estos valores, sin embargo, al remover una fila entera con un entrenador “NaN” puede también remover el entrenador del equipo rival, este procedimiento estaría removiendo información valiosa. Para reducir el número de datos faltantes, razón por la cual se completaron los datos “NaN”, con el entrenador anterior de ese mismo equipo, este proceso se realizó recorriendo el Dataset de forma inversa, y rellenando los espacios vacíos con el ultimo entrenador que el club tuvo. Realizar esta operación no creó consecuencias negativas ya que, si un entrenador es despedido o abandona el equipo, dejara su estilo de juego en la plantilla que se conservara mientras

un nuevo entrenador llega (esto en caso de que no sea reemplazado de forma inmediata).

El porcentaje de datos perdidos en las columnas referentes a los entrenadores se ha reducido a un 4%, aun se debe remover del DataFrame, las filas que no poseen el registro de goles debido a que no se han jugado esos encuentros, por lo tanto, se procedió a eliminar estos datos.

El resultado final es un nuevo Dataset llamado “europa_resultados_limpios” el cual contiene 187907 filas por 11 columnas.

Enseguida se completan los valores “NaN” faltantes con el String “sin entrenador” para el caso de las columnas “entrenador_local” y “entrenador_visitante” y se completan los valores faltantes con el String “sin país” para las columnas “pais_entrenador_local” y “pais_entrenador_visitante”, este proceso se realiza con el fin de no dejar valores “NaN”, pues no es aconsejable emplear transformaciones o entrenar modelos utilizando Dataset con este tipo de datos ya que puede presentar errores.

El Dataset no posee una variable que indique el número de puntos actuales de los equipos que se van a enfrentar, por lo tanto, se implementó un código para recorrer el Dataset por equipo, y luego por partido jugado, para acumular los puntos actuales y asignarlos como dos nuevas variables al Dataset, una columna para puntos del equipo local y otras para el equipo visitante, estas nuevas columnas tienen los nombre de “puntos_actuales_local” y “puntos_actuales_visitante”.

- ***Transformación de Dataframe basado en el cambio de entrenador***

Una vez está limpio el Dataset, se realizó una transformación mediante código para recorrerlo de equipo en equipo detectando los cambios de entrenador, cuando el código detecta el cambio de entrenador, se obtienen diversos datos que se describirán a continuación.

De esta manera se genera un nuevo Dataset llamado “europa_promedio_goles” donde cada instancia contiene diversas variables en el momento del cambio de entrenador.

Se repite el proceso para los Datasets referentes a las ligas europeas femeninas y para las ligas latinoamericanas, resultando en los siguientes tres Datasets:

- europa_promedio_goles (6977 filas por 118 columnas).
- europa_femenina_promedio_goles (380 filas por 18 columnas).
- latinoamerica_promedio_goles (2167 filas por 18 columnas).

- ***Fusión de Datasets promedio de goles con Dataset de La FiFA***

Para mejorar el Dataset “europa_promedio_goles” el cual contiene el mayor número de estancias, se realiza proceso de combinación de el Dataset “europa_promedio_goles” para agregar algunos atributos como lo son el ataque y la defensa que posee el equipo según el Dataset de datos generado por La FIFA, para realizar este proceso, se crea un código en Python que recorre el Dataset “europa_promedio_goles” y para cada fila recorre el Dataset de LA FIFA en busca del equipo que acaba de cambiar de entrenador teniendo en cuenta la fecha en la que se dio el cambio para poder asignar las nuevas características.

El Dataset de La FIFA solo tiene datos hasta el año 2005, así que desde el año 2005 hasta el 2021, se asignaran las características a las filas según el año correspondiente, mientras que para las filas con fechas inferiores a el año 2005 solo se les atribuirá las características de este año, (Recordar que en el Dataset “europa_promedio_goles”, cada fila representa un evento de cambio de entrenador).

Originalmente el Dataset “europa_promedio_goles”, posee 6977 filas por 18 columnas, luego de ejecutar la combinación con el Dataset de La FIFA se va a ver reducido pues no todos los equipos que se encuentran en el Dataset de Europa se encuentran en el Dataset de La FIFA.

Después de la combinación, el Dataset “promedio_goles_europa” contiene 3 variables nuevas que indican el promedio de ataque, medio y defensa del equipo correspondiente a esa fila, 14% de los datos aparecen como datos “NaN”, haciendo referencia a los equipos faltantes, estas filas se deben remover del Dataset ya que no aportarían nada a los modelos de aprendizaje.

El nuevo Dataset contiene las siguientes dimensiones:

- europa_promedio_goles_fifa (6025 filas por 21 columnas).

El anterior Dataset contiene todas las variables del Dataset “europa_promedio_goles”, más las variables “ataque”, “medio” y “defensa” provenientes del Dataset “fifa_caracteristicas_equipos”.

A el Datasets “europa_femenina_promedio_goles” no se le realizó este procedimiento dado que solo hay información de ligas femeninas en el Dataset de La FIFA desde el año 2016 lo que proporciona muy poca información y para el Dataset “latinoamerica_promedio_goles” no se le aplico el proceso, ya que hay equipos de la liga colombiana de los que no se tiene registro hasta el año 2015, por lo que son muy pocos datos para realizar la fusión (fifaindex, s.f.).

4. Fase IV. Modeling

Esta sección se dividió en dos partes, la primera parte se enfocó en el análisis estadístico de los Datasets referentes a Europa, Europa femenina y Latinoamérica, donde se utilizó como herramienta la regresión lineal para detectar con mayor facilidad el comportamiento de las variables de los Datasets, como los tres Datasets poseen las mismas variables, se realizó una comparación de medias y en la siguiente fase (Fase V. Evaluation), se pueden observar los resultados y comparaciones del modelo de regresión lineal.

La segunda parte del Dataset está enfocada en la construcción de 4 modelos de aprendizaje supervisado los cuales tienen como meta intentar predecir el promedio de puntos o promedio de goles que un equipo puede llegar a generar después de que cambia de entrenador, para esta sección se utiliza el Dataset de Europa el cual posee el mayor número de datos, combinado con el Dataset extraído de la FIFA construido en la fase III. Para la construcción de estos modelos se utilizan como herramientas las técnicas Decision Tree, Random Forest, AdaBoost y el GradientBoosting, con el fin de elegir la mejor técnica de predicción al finalizar la fase de modelado.

Parte 1: Análisis estadístico de las ligas masculinas y femeninas

- **Análisis de Dataset, comparación de medias**

Para el análisis de las siguientes variables, solo se mostrarán los aspectos más importantes de las variables con datos sobre porcentajes y promedios de los 5 encuentros anteriores y posteriores al cambio de entrenador.

	Europa		Europa F		Latinoamérica	
	Mediana	Media	Mediana	Media	Mediana	Media
promedio_goles_hechos_antes..	1.2	1.23	1.29	1.53	1.0	1.08
promedio_goles_hechos_despues..	1.2	1.29	1.4	1.62	1.0	1.14
promedio_goles_recibidos_antes..	1.6	1.66	1.6	1.85	1.4	1.43
promedio_goles_recibidos_despues..	1.4	1.46	1.5	1.68	1.2	1.26
puntos_hechos_antes..	5.0	5.35	6.0	6.3	5.0	5.37
puntos_hechos_despues..	6.0	5.80	6.0	6.56	6.0	5.53
porcentaje_victorias_despues..	0.33	0.32	0.4	0.4	0.4	0.32
porcentaje_victorias_antes..	0.2	0.27	0.4	0.36	0.2	0.26
porcentaje_empates_despues..	0.2	0.26	0.2	0.17	0.2	0.28
porcentaje_empates_antes..	0.2	0.24	0.2	0.16	0.2	0.27
porcentaje_derrotas_despues..	0.4	0.40	0.4	0.41	0.4	0.38
porcentaje_derrotas_antes..	0.4	0.47	0.4	0.46	0.4	0.45

Tabla 1. Resumen de medianas y medias de las variables para los Dataset "europa_promedio_goles", "europa_promedio_goles_femenina" y "latinoamerica_promedio_goles". Elaboración propia.

	Europa		Europa F		Latinoamérica	
	Mediana	Media	Mediana	Media	Mediana	Media
promedio_goles_hechos_despues..	0%	4.8%	8.5%	5.8%	0%	5.5%
promedio_goles_recibidos_despues..	-14.2%	-13.6%	-6.6%	-10.1%	-16.6%	-13.4%
puntos_hechos_despues..	20%	8.4%	0%	4.1%	20%	2.97%
porcentaje_victorias_despues..	65%	18.5%	0%	11.1%	100%	23%
porcentaje_empates_despues..	0%	8.3%	0%	6.2%	0%	3.7%
porcentaje_derrotas_despues..	0%	-17%	0%	-12.1%	0%	-18.4%

Tabla 2. Porcentaje de aumentos y decrementos de medias y medianas después de cambio de entrenador. Elaboración propia.

En la tabla 1, se pueden observar las medianas y medias para cada Dataset de las variables con los datos referentes a los 5 encuentros antes y después del cambio de entrenador.

La variable “promedio_goles_hechos_antes_de_cambio_entrenador_1_5”, para el Dataset “europa_promedio_goles”, tuvo una mediana de 1.2 y una media de 1.23, estos valores tienen una relación con la siguiente variable que se analizara, la cual representa los goles hechos después del cambio de entrenador. Para el Dataset “europa_femenina_promedio_goles”, la mediana es de 1.29 y la media es de 1,53, estos valores también poseen una relación con la variable “promedio_goles_hechos_despues_de_cambio_entrenador_1_5”, para el Dataset “latinoamerica_promedio_goles”, se obtiene un valor de 1 para la media y 1,08 para la mediana.

Como se puede observar en la tabla 1 la variable “promedio_goles_hechos_despues_de_cambio_entrenador”, la mediana es la misma que se obtuvo en la variable “promedio_goles_hechos_antes_de_cambio_entrenador_1_5”, y la media aumenta ligeramente, para el caso de Europa, aumento de 1.23 a 1.29, para Europa femenina aumento de 1.53 a 1.62, y para Latinoamérica aumento de 1.08 a 1.14, en los tres casos, aumento la media.

Para la variable “promedio_goles_recibidos_antes_de_cambio_entrenador_1_5”, en el caso de Europa, la mediana es de 1.6 y la media es de 1.66, para Europa femenina la mediana es de 1.6 y la media de 1.85 y para Latinoamérica la mediana es de 1.4 y la media es de 1.43, esta variable tiene resultados similares con su variable compañera directa, la cual es “promedio_goles_recibidos_despues_de_cambio_entrenador_1_5”.

Para la variable “Promedio_goles_recibidos_despues_de_cambio_entrenador_1_5” del Dataset de promedio de goles de Europa, la mediana fue de 1.4, y la media de 1.46, para Europa femenina la mediana fue de 1.5 y la media de 1.68 y para Latinoamérica la mediana fue de 1.2 y la media de 1.26, como se puede observar tanto la mediana como la media se han reducido en los tres Dataset.

Para el caso de Europa, la variable “puntos_hechos_antes_de_cambio_entrenador” posee el valor de mediana de 5 y el valor de media de 5.35, para Europa femenina la mediana es de 6 y la media de 6.3, para Latinoamerica la mediana es de 5 y la media de 5.37 , los valores obtenidos son inferiores en comparación la variable “puntos_hechos_despues_de_cambio_entrenador_1_5”, ya que esta variable presenta aumento en todas las medias e igualdad en algunas las medianas, pero nunca un decremento ni en la mediana y en la media, para el caso de Europa la mediana fue de 6 y la media de 5.8, para Europa femenina la mediana fue de 6 y la media fue de 6.56, para Latinoamérica la mediana fue de 6 y la media de 5.53.

La mediana para el porcentaje de victorias en los 5 encuentros después del cambio de entrenador para Europa es de 0.33 y la media es de 0.32, para Europa Femenina la mediana es de 0.4 y la media de 0.4, para Latinoamérica la mediana es de 0.4 y la media de 0.32, similar a los casos anteriores, la variable “porcentaje_vistorias_antes_de_cambio_entrenador_1_5” presenta valores muy cercanos, levemente reducidos para el caso de las medias y prácticamente iguales para el caso de las medianas, este efecto se repite para las variables referentes a las victorias y derrotas siendo esta última la única variable donde el efecto es inverso, es decir que los valores de media disminuyen levemente después de realizar el cambio de entrenador.

Como se observa en la tabla 1, los valores de las medianas se mantuvieron prácticamente intactos en todas las variables, mientras que las medias de todas las variables aumentaban para el caso del promedio goles hechos, puntos conseguidos promedio de victorias y promedio de empates, después del cambio de entrenador, mientras que la media para las variables que se refieren a sucesos negativos para un equipo como el promedio de goles recibidos, y el promedio de derrotas se redujo levemente. Sin embargo, estos cambios no reflejan cambios relevantes y se puede considerar que no hay un aumento significativo de los datos obtenidos en los encuentros anteriores y posteriores al momento de cambiar de entrenador.

- **Análisis de Dataset, comparación de desviación estándar**

En la siguiente tabla se puede observar la desviación estándar y el coeficiente de variación para las variables de los tres Datasets.

	Europa		Europa F		Latinoamérica	
	Std	CV	Std	CV	Std	CV
promedio_goles_hechos_antes..	0.66	0.54	0.96	0.62	0.55	0.50
promedio_goles_hechos_despues..	0.73	0.56	1.19	0.73	0.64	0.56
promedio_goles_recibidos_antes..	0.68	0.41	1.08	0.58	0.58	0.40
promedio_goles_recibidos_despues..	0.76	0.52	1.14	0.68	0.69	0.54
puntos_hechos_antes..	3.18	0.59	4.50	0.71	2.97	0.55
puntos_hechos_despues..	3.44	0.59	4.73	0.72	3.41	0.61
porcentaje_victorias_despues..	0.25	0.78	0.33	0.82	0.26	0.81
porcentaje_victorias_antes..	0.22	0.81	0.31	0.86	0.20	0.78
porcentaje_empates_despues..	0.23	0.87	0.22	1.23	0.25	0.87
porcentaje_empates_antes..	0.19	0.80	0.20	1.24	0.19	0.72
porcentaje_derrotas_despues..	0.27	0.68	0.33	0.79	0.28	0.74
porcentaje_derrotas_antes..	0.24	0.51	0.31	0.68	0.23	0.52

Tabla 3. Resumen de desviación estándar y coeficiente de variación de las variables para los Dataset "europa_promedio_goles", "europa_promedio_goles_femenina" y "latinoamerica_promedio_goles". Elaboración propia.

Similar al caso de las medias de los tres Datasets, la desviación estándar se incrementó en todas las variables relacionadas posterior al cambio de entrenador, en algunas variables se incrementa más que en otras, por ejemplo la variable "promedio_goles_hechos_antes_de_cambio_entrenador_1_5", se incrementó ligeramente para el caso de Europa y Latinoamérica de 0.66 a 0.73 y de 0.55 a 0.64 respectivamente, mientras que para Europa Femenina el incremento fue mayor, pasando de 0.96 a 1.19; En las ligas femeninas la desviación con respecto a la media es superior para la variable en cuestión.

La desviación estándar en las variables del Dataset Europa Femenina poseen una mayor magnitud que las variables para los otros dos Datasets, excepto en la variable "porcentaje_empates_antes_de_cambio_entrenador", dando a entender que los datos referentes a las ligas femeninas europeas poseen una mayor dispersión con respecto a la media, de hecho la media de las ligas femeninas también posee valores más altos o iguales comparados con los datos de las ligas masculinas Europeas y Latinoamericanas.

El coeficiente de variación para las características es alto en todos los Datasets, en su mayoría superando el 0.5 lo cual indica que la varianza es alta, siendo superior en el Dataset de las ligas femeninas europeas, esta particularidad está relacionada con la alta dispersión que poseen los datos.

Los datos más relevantes que se pueden tomar son las medias y medianas de los Datasets, ya que enseñan si en una variable especifica se dieron cambios después del cambio de entrenador, para este caso se puede observar en la tabla 1 que los cambios en las medias y medianas de las variables antes y después del cambio de entrenador son apenas notables prácticamente no existió cambio apreciable; El coeficientes de variación, en la tabla 2 indica que todas las variables sufren de alta variabilidad, sobre todo en las ligas femeninas, también destacar que posterior al cambio de entrenador, las variables aumentan el coeficiente de variación.

- **Agrupamiento de datos por medias de equipos.**

Para cada uno de los Dataset se realiza un agrupamiento de medias para cada uno de los equipos, se genera un nuevo DataFrame con la acumulación de dichos valores, este proceso se realizó recorriendo cada Dataset acumulando la media de cada una de las variables, evitando variables como la fecha y la jornada y los puntos actuales ya que es una acumulación de medias donde no tendría sentido incluir la media de la fecha, de esta manera cada fila del DataFrame representara a un club deportivo y las medias de diferentes variables que se presentaran a continuación:

Las características de los 3 nuevos DataFrame son las siguientes:

- **equipo**: Nombre del equipo
- **m_goles_hechos_antes**: Media de goles hechos antes cambio del entrenador.
- **m_goles_hechos_despues**: Media de goles hechos después del cambio de entrenador.
- **m_goles_recibidos_antes:** Media de goles recibidos antes del cambio de entrenador.
- **m_goles_recibidos_despues:** Media de goles recibidos después del cambio de entrenador.
- **m_puntos_hechos_antes:** Media de puntos hechos antes del cambio de entrenador.
- **m_puntos_hechos_despues:** Media de goles hechos después del cambio de entrenador.
- **m_victorias_antes:** Media de victorias antes del cambio de entrenador.
- **m_victorias_despues:** Media de victorias después del cambio de entrenador.

- Correlación entre variables

Las variables de los nuevos DataFrames contienen los datos sobre las medias por equipo individual, en seguida se generó una matriz de correlación entre las variables de los tres DataFrames con el fin de observar justamente la correlación que existe entre las características.

Las variables que contienen datos con información referente a un club luego del cambio de entrenador, es decir las variables con la palabra “después”, poseen una correlación fuerte con otras variables del mismo tipo, por ejemplo la variable “m_goles_hechos_despues” posee una correlación alta (mayor a 0.7) con las variables “m_victorias_despues” y “m_puntos_hechos_despues”, con la única variable que no presenta una correlación fuerte e incluso negativa es con la variable “m_goles_recibidos_despues”, sin embargo la variable en cuestión posee una correlación mediana (mayor a 0.4) con las variables con datos previos al cambio de entrenador, estas variables tienen la palabra “antes” en su nombre, para este caso, la variable en cuestión tiene correlación media con las variables “m_victorias_antes” y “m_puntos_hechos_antes”, pero una poca correlación con la variable

“m_goles_recibidos_antes”, este mismo patrón se repite con las demás variables con la palabra “después” referentes información posterior al cambio de entrenador, excepto para el caso de la variable “m_goles_recibidos_despues”, ya que las correlaciones que posee con las demás variables son negativas aunque con magnitudes similares.

Para el caso de las ligas latinoamericanas y las ligas femeninas las correlaciones entre las variables son similares a las correlaciones del DataFrame Europa, salvo por pequeñas diferencias como la correlación de la variable “m_goles_hechos_antes” y “m_goles_hechos_despues”, la cual es prácticamente nula en las ligas latinoamericanas, el resto de variables se mantienen iguales.

Las correlaciones entre las variables con información de los encuentros posteriores al cambio de entrenador son altas, esto es un hecho que se esperaba ya que si un equipo anota muchos goles en los partidos subsecuentes al cambio de entrenador es muy probable que también gane dichos partidos, de igual manera, si un equipo recibe pocas anotaciones de gol en su portería, es probable que su índice de victoria sea alto, pues tienen una correlación negativa.

- **Regresión lineal para detectar linealidad en variables antes y después del cambio de entrenador.**

La regresión lineal es útil para observar la linealidad si es que existe entre las variables correspondientes, como se puede observar, los tres Datasets, poseen las mismas variables con diferentes valores, cada variable posee una compañera, una contiene información sobre la media de goles, victorias o puntos previo al cambio de entrenador, mientras que la otra contiene la misma información después de ejecutar el cambio de entrenador. El objetivo de la regresión en este caso es indagar entre cada pareja de variables para tratar de obtener y visualizar la linealidad mediante los coeficientes que generen los modelos de regresión, primero se aplicó la regresión a las variables del Dataframe Europa luego para el Dataframe Latinoamérica y finalmente para el Dataframe de las ligas femeninas.

Antes de implementar los modelos de regresión lineal, se eliminaron los Outliers del Dataset, este proceso se realizó en este punto ya que, si se eliminan los Outliers de todas las variables, se reduciría el número de muestras, mientras que si se realiza teniendo en cuenta solo las dos variables con las que se construirá el modelo, se contaran con más datos. Para eliminar las filas con Outliers en estas variables, se calculan los bigotes superior e inferior basados en el rango intercuartílico, para calcular el bigote inferior se le resta al primer cuartil (Q1) de cada variable, el rango intercuartílico de cada variable (Q3-Q1) multiplicado por 1.5, y para calcular el bigote superior se le suma al tercer cuartil (Q3), el rango intercuartílico de cada variable, multiplicado por 1.5, este proceso se aplica para cada par de variables.

Una vez calculados los bigotes superiores e inferiores de cada variable en los tres Dataframes, se eliminan los valores Outliers, de las dos variables que se van a utilizar en la regresión lineal, en la tabla 3 se encuentran las variables que se utilizaron en cada regresión lineal, el valor de los bigotes para cada variable y la cantidad de muestras que

se utilizaron para el entrenamiento y el testeo, de color gris se encuentran subrayados los modelos para las ligas Europeas masculinas, de color verde se encuentran subrayados los modelos para las ligas Latinoamericanas y de color naranja se encuentran subrayados los modelos para las ligas femeninas.

Modelo	Variable	Tipo de variable	Bigote inferior	Bigote superior	Número de variables para Train	Número de variables para Test
1	m_goles_hechos_antes	Independiente	0.37	1.86	304	76
	m_goles_hechos_despues	Dependiente	0.29	2.02		
2	m_victorias_antes	Independiente	0.02	0.46	291	73
	m_victorias_despues	Dependiente	-0.01	0.56		
3	m_goles_recibidos_antes	Independiente	0.88	2.60	300	75
	m_goles_recibidos_despues	Dependiente	0.67	2.47		
4	m_puntos_hechos_antes	Independiente	1.26	8.39	298	75
	m_puntos_hechos_despues	Dependiente	0.86	9.22		
1	m_goles_hechos_antes	Independiente	0.46	1.60	93	24
	m_goles_hechos_despues	Dependiente	0.36	1.79		
2	m_victorias_antes	Independiente	0.01	0.50	92	24
	m_victorias_despues	Dependiente	0.04	0.55		
3	m_goles_recibidos_antes	Independiente	0.70	2.26	94	24
	m_goles_recibidos_despues	Dependiente	0.58	2.01		
4	m_puntos_hechos_antes	Independiente	1.51	8.69	89	23
	m_puntos_hechos_despues	Dependiente	2.01	8.46		
1	m_goles_hechos_antes	Independiente	-0.52	3.29	108	28
	m_goles_hechos_despues	Dependiente	-0.5	3.5		
2	m_victorias_antes	Independiente	-0.42	1.13	92	24
	m_victorias_despues	Dependiente	-0.39	1.2		
3	m_goles_recibidos_antes	Independiente	-1.25	4.75	113	29
	m_goles_recibidos_despues	Dependiente	-0.80	4.0		
4	m_puntos_hechos_antes	Independiente	-6.0	18.0	116	29
	m_puntos_hechos_despues	Dependiente	-4.87	-17.72		

Tabla 4 . Número de muestras usadas y posición de bigotes de las variables de los Dataset "europa_promedio_goles", "europa_promedio_goles_femenina" y "latinoamerica_promedio_goles". Elaboración propia.

- ***Parte 2: Aplicación de modelos de aprendizaje supervisado***

En esta sección se realizaron varios modelos de aprendizaje supervisado utilizando el Datasets "europa_promedio_goles_fifa" utilizando las variables seleccionadas en la fase III, con el fin de tratar de predecir el promedio de goles que un equipo marcara en los 5 partidos posteriores a realizar un cambio de entrenador deportivo y el número de puntos que un equipo puede generar en los 5 partidos siguientes al cambio de entrenador. De esta manera, un equipo puede realizar una predicción y determinar si es un buen momento cambiar el entrenador o no, basado en el promedio de goles antes del cambio, la jornada actual y el número de puntos antes del cambio etc.

Se aplicaron las técnicas de aprendizaje supervisado “RandomForests”, “Decision Tree”, “AdaBoost” y “GradientBoosting”, para tratar de predecir algunas variables objetivo relacionadas con el promedio de goles hechos o recibidos después del cambio de entrenador de un equipo, también realizó la predicción del promedio de puntos que un equipo puede llegar a conseguir después del cambio de entrenador, se utilizaron las 4 técnicas de aprendizaje para luego comparar los resultados obtenidos.

Las 4 técnicas de modelado se han seleccionado debido a que los métodos que utilizan árboles o conjuntos de árboles poseen cualidades actas para la predicción en problemas de clasificación y regresión, sobre todo los “Ensembles” o conjuntos de modelos ya que son resistentes a los valores Outliers y tienen buena escalabilidad lo que permite ser aplicados a conjuntos de datos con muchas observaciones. (Amat, 2020).

- **Modelos de aprendizaje aplicados a la variable “puntos_hechos_despues_de_cambio_entrenador_1_5”**

  - Preproceso antes de aplicar modelos

Antes de implementar alguno de los modelos de aprendizaje supervisado, se deben eliminar los Outliers del Dataset, este proceso se realiza en este punto ya que, si se eliminan los Outliers de todas las variables, se reduciría el número de muestras, mientras que si se realiza una vez se han seleccionado las características, solo se borraran las filas que contengan Outliers para las variables elegidas, para ello se calcularon nuevamente los rangos intercuartílicos para obtener el valor de los bigotes superiores e inferiores.

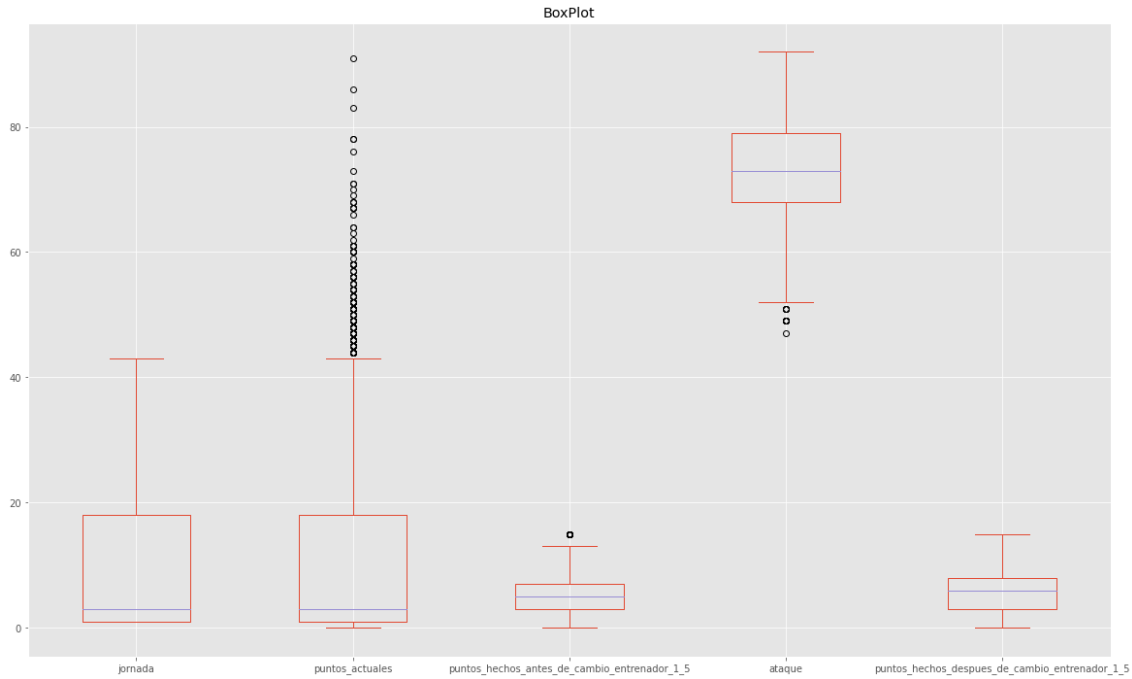

Ilustración 1. Boxplot de las variables seleccionadas (Variable objetivo: puntos_hechos_despues_de_cambio_entrenador_1_5). Fuente: Elaboración propia.

En la gráfica se pueden observar los Outliers de las variables seleccionadas y la variable objetivo, para eliminar las filas con este tipo de datos basados en estas variables, se calcularon los bigotes superior e inferior basados en el rango intercuartílico, para calcular el bigote inferior se le resto al primer cuartil (Q1) de cada variable el rango intercuartílico de cada variable (Q3-Q1) multiplicado por 1.5, y para calcular el bigote superior se le sumo al tercer cuartil (Q3), el rango intercuartílico de cada variable, multiplicado por 1.5.

Una vez limpio el Dataset de la mayoría de Outliers se normalizan los datos en un rango de valores entre 0 y 1 con el fin de que las variables que posean un rango de números elevados no obtengan prioridad en el modelo respecto a otras variables.

Luego de normalizar los datos se divide el Dataset en dos partes, una parte contiene el 80% de los datos para entrenar el modelo y la otra parte el 20% de los datos para realizar el testeo del modelo, esta división se realiza de forma aleatoria.

La variable que se tratara de predecir es la variable "puntos_hechos_despues_de_cambio_entrenador_1_5", la cual representa la sumatoria de puntos hechos en los 5 encuentros después de que un equipo realiza un cambio de entrenador.

- RandomForests para predicción de puntos obtenidos después del cambio de entrenador

El primer modelo que se utilizó para intentar realizar la predicción de la variable relacionada con el promedio de goles hechos luego del reemplazo de un entrenador fue el RandomForests, sin embargo, antes de aplicar el modelo, se realizó el proceso de Cross Validation para determinar el mejor número de estimador, iterando entre varias potencias del 2.

Modelo	RandomForests
Variable objetivo	puntos_hechos_despues_de_cambio_entrenador_1_5
Mejor MAE	2.52
n.º Estimadores	1024

Tabla 5. Número de estimadores óptimo para el modelo RandomForest (variable: puntos_hechos_despues_de_cambio_entrenador_1_5). Elaboración propia.

Se determina el número de estimador con el MAE más bajo, en este caso el mejor resultado, es el número 1024.

Una vez identificado el mejor número de estimadores, se construyó el modelo. Se entreno con la división creada previamente que contiene el 80% de los datos y se probó con la división del 20%, para la construcción del modelo se utilizaron las variables seleccionadas en la fase III.

Características seleccionadas:

- jornada
- puntos_actuales
- puntos_hechos_antes_de_cambio_entrendor_1_5
- ataque

Variable objetivo:

- puntos_hechos_despues_de_cambio_entrendor_1_5

Hiperparámetros del modelo:

- n_estimators: 1024
- criterion: “mae”
- max_depth: Default (None)
- random_state: 0

Una vez entrenado el modelo de regresión, este procede a ser probado con la división de testeo. (Pedregosa, 2011)

- DecisionTree para predicción de puntos obtenidos después de cambio de entrenador

Se continua con el siguiente modelo el cual es un árbol de decisión para intentar predecir el número de puntos obtenidos después del cambio de entrenador, se utilizan los mismos datos normalizados del anterior modelo, primero se realiza el proceso de “Cross Validation Analysis” para determinar la máxima profundidad del árbol que se debería usar para conseguir el menor error absoluto medio.

Modelo	DecisionTree
Variable objetivo	puntos_hechos_despues_de_cambio_entrenador_1_5
Mejor MAE	2.49
Máxima profundidad	5

Tabla 6. Número de máxima profundidad para el modelo RandomForest (variable: puntos_hechos_despues_de_cambio_entrenador_1_5). Elaboración propia.

Una vez obtenido el "max_depth" (máxima profundidad del árbol), se procedió a implementar el modelo DecisionTreeRegressor de la librería sickit-learn (Pedregosa, 2011).

Características seleccionadas:

- jornada.
- puntos_actuales.
- porcentaje_victorias_antes_de_cambio_entrenador_1_5
- puntos_hechos_antes_de_cambio_entrendor_1_5

Variable objetivo:

- puntos_hechos_despues_de_cambio_entrendor_1_5

Hiperparámetros del modelo:

- max_depth: 5
- criterion: "mae"

El modelo de árbol de decisión se construyó con los hiperparámetros seleccionados y se entrenó con la división de entreno que contiene el 80% de los datos, posteriormente se prueba con los datos de test.

- **AdaBoost para predicción de puntos obtenidos después de cambio de entrenador**

A continuación, se implementa la técnica "AdaBoost", se realiza el proceso de Cross Validation para determinar el número de estimadores que en este caso representa el número iteraciones que el algoritmo de AdaBoost realizará antes de finaliza el refuerzo.

Modelo	AdaBoost
Variable objetivo	puntos_hechos_despues_de_cambio_entrenador_1_5
Mejor MAE	2.46
n.º Estimadores	16

Tabla 7. Número de estimadores óptimo para el modelo AdaBoost (variable: puntos_hechos_despues_de_cambio_entrenador_1_5). Elaboración propia.

Se utilizará el "max_depth" obtenido para el árbol de decisión, y el criterio "MAE", junto con el número de estimadores obtenido en el Cross Validation como parámetros para la implementación del modelo AdaBoost.

Características seleccionadas:

- jornada
- puntos_actuales
- puntos_hechos_antes_de_cambio_entrendor_1_5
- ataque

Variable objetivo:

- puntos_hechos_despues_de_cambio_entrendor_1_5

Hiperparámetros del modelo:

- n_estimators: 16
- learning_rate: Default (0.01)
- loss: Default (linear)
- random_state: 0
- criterion_ “mae” (DecisionTree)
- max_depth: 5 (DecisionTree)

Se construye el modelo utilizando los Hiperparámetros y se entrena con los datos de entrenamiento para posteriormente probarlos (Pedregosa, 2011).

- GradientBoosting para predicción de puntos después de cambio de entrenador.

Se realizó en proceso de Cross Validadion para determinar el mejor número de estimador (para el caso de GradientBoosting, el número de estimador indica el número de etapas de refuerzo, entre más grande el número puede generar mejores resultados ya que este modelo es resistente al Overfitting (Pedregosa, 2011). El estimador se iterará entre varias potencias del 2.

Modelo	GradientBoosting
Variable objetivo	puntos_hechos_despues_de_cambio_entrenador_1_5
Mejor MAE	2.46
n.º Estimadores	512

Tabla 8. Número de estimadores óptimo para el modelo GradientBoosting (variable: puntos_hechos_despues_de_cambio_entrenador_1_5). Elaboración propia.

Se obtiene el mejor valor de número de estimador para obtener el menor MAE.

Características seleccionadas:

- jornada.
- puntos_actuales.
- puntos_hechos_antes_de_cambio_entrendor_1_5

- ataque

Variable objetivo:

- puntos_hechos_despues_de_cambio_entrendor_1_5

Hiperparámetros del modelo:

- n_estimators: 512
- criterion: “mae”
- learning_rate: 0.01
- loss: “absolute_error”
- random_state: 0

Se construye el modelo de regresión y se entrena con el 80% de los datos.

- Modelos de aprendizaje aplicados a la variable “promedio_goles_hechos_despues_de_cambio_entrenador_1_5”

Se han aplicado los 4 modelos para tratar de predecir la variable “puntos_hechos_antes_de_cambio_entrenador_1_5”, los resultados se pueden observar en la siguiente fase (fase V), enseguida se aplicarán las 4 técnicas de aprendizaje supervisado para intentar predecir la variable llamada “promedio_goles_hechos_despues_de_cambio_entrendor_1_5” utilizando otra selección de características, se realiza el mismo proceso de eliminación de Outliers utilizando anteriormente pero con las nuevas características.

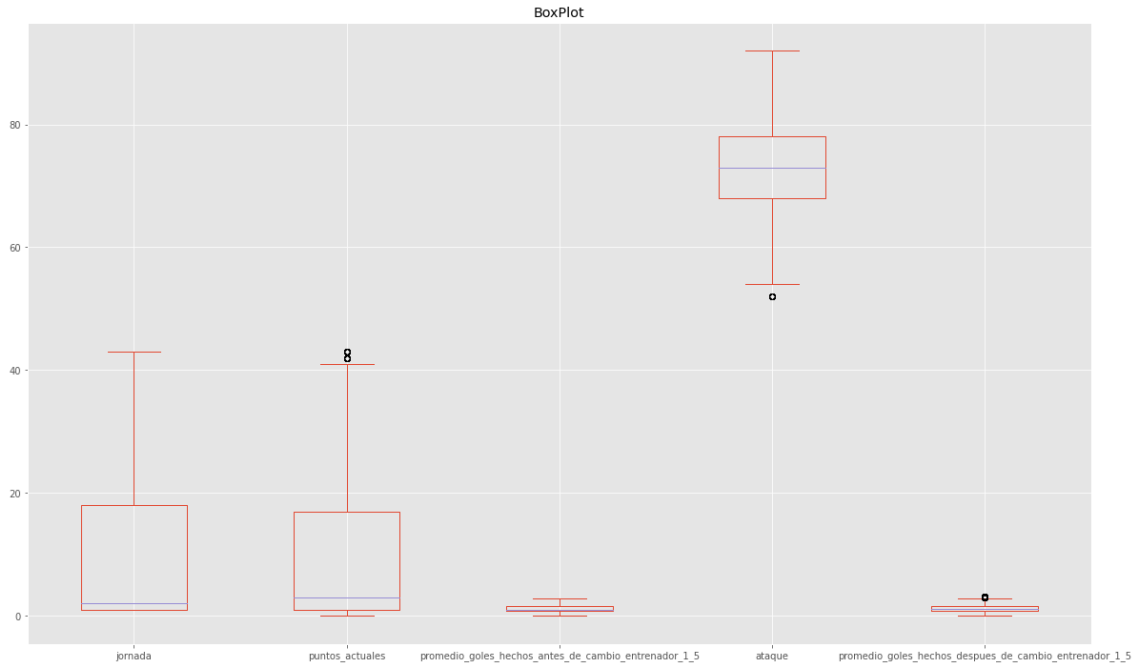

Ilustración 2. Boxplot de las variables seleccionadas sin outliers. Fuente: Elaboración propia.

- RandomForests para predicción de goles hechos después del cambio de entrenador

Se utilizó nuevamente la técnica RandomForests, en el mismo Dataset pero para predecir la variable "promedio_goles_hechos_despues_de_cambio_entrenador_1_5", también cambiaran las características seleccionadas para realizar el entrenamiento del modelo, como en el anterior RandomForests, se utilizara el Cross Validation para seleccionar el mejor estimador utilizando varias potencias del 2.

Modelo	RandomForests
Variable objetivo	promedio_goles_hechos_despues_de_cambio_entrenador_1_5
Mejor MAE	0.51
n.º Estimadores	1024

Tabla 9. Número de estimadores óptimo para el modelo RandomForests (Variable: promedio_goles_hechos_despues_de_cambio_entrenador_1_5). Elaboración propia.

Se obtuvo un MAE de 0.54 con un número de estimador de 1024.

Características seleccionadas:

- jornada
- puntos_actuales
- promedio_goles_hechos_antes_de_cambio_entrenador_1_5
- ataque

Variable objetivo:

- promedio_de_goles_hechos_despues_de_cambio_entrendor_1_5

Hiperparámetros del modelo:

- n_estimators: 1024
- criterion: "mae"
- max_depth: Default (None)
- random_state: 0

Se entreno el modelo regresión utilizando los hiperparámetros anteriores y luego se prueba con el conjunto de test que contiene el 20% de los datos.

- DecisionTree para Predicción goles hechos después de cambio de entrenador

Nuevamente se utilizó la técnica Cross Validation para definir la profundidad del árbol adecuada, se realizaron iteraciones probando valores para la profundidad del árbol entre un rango de 2 a 50.

Modelo	DecisionTree
Variable objetivo	Promedio_goles_hechos_despues_de_cambio_entrenador_1_5
Mejor MAE	0.48
Máxima profundidad	4

Tabla 10. Número de máxima profundidad para el modelo DecisionTree (Variable: promedio_goles_hechos_despues_de_cambio_entrenador_1_5). Elaboración propia.

Según el Cross Validation el mejor número de profundidad del árbol es de 4.

Características seleccionadas:

- jornada
- puntos_actuales
- promedio_goles_hechos_antes_de_cambio_entrenador_1_5
- promedio_goles_recividos_antes_cambio_entrenador_1_5

Variable objetivo:

- promedio_de_goles_hechos_despues_de_cambio_entrendor_1_5

Hiperparámetros del modelo:

- max_depth: 4
- criterion: “mae”

Se construye el árbol de decisión con los Hiperparámetros seleccionados.

- AdaBoost para Predicción goles hechos después de cambio de entrenador.

Se empleó la técnica Cross Validation nuevamente para obtener el mejor número de estimadores.

Modelo	AdaBoosting

Variable objetivo	Promedio_goles_hechos_despues_de_cambio_entrenador_1_5
Mejor MAE	0.48
n.º Estimadores	8

Tabla 11. Número de estimadores óptimo para el modelo AdaBoosting (Variable: promedio_goles_hechos_despues_de_cambio_entrenador_1_5). Elaboración propia.

Se obtiene el número de estimadores con el valor de 8.

Características seleccionadas:

- jornada
- puntos_actuales
- promedio_goles_hechos_antes_de_cambio_entrenador_1_5
- ataque

Variable objetivo:

- promedio_de_goles_hechos_despues_de_cambio_entrendor_1_5

Hiperparámetros del modelo:

- n_estimators: 8
- learning_rate: Default (0.01)
- loss: Default ("linear")
- random_state: 0
- criterion: "mae" (DecisionTree)
- max_depth: 4 (DecisionTree)

Se construyo el modelo utilizando los hiperparámetros en el conjunto de datos de entrenamiento.

- GradientBoosting para Predicción goles hechos después de cambio de entrenador.

Se utiliza el Cross Validadion con varias potencias del 2 para determinar el mejor número de estimador.

Modelo	GradientBoosting
Variable objetivo	promedio_goles_hechos_despues_de_cambio_entrenador_1_5
Mejor MAE	0.48
n.º Estimadores	128

Tabla 12. Número de estimadores óptimo para el modelo GradientBoosting (Variable objetivo: promedio_goles_hechos_despues_de_cambio_entrenador_1_5). Elaboración propia.

Obtenemos el mejor valor de número de estimador 1024.

Características seleccionadas:

- jornada
- puntos_actuales
- promedio_goles_hechos_antes_de_cambio_entrenador_1_5
- ataque

Variable objetivo:

- promedio_de_goles_hechos_despues_de_cambio_entrendor_1_5

Hiperparámetros del modelo:

- n_estimators: 128
- criterion: “mae”
- learning_rate: 0.01
- loss: “absolute_error”
- random_state: 0

Se construye el modelo utilizando los hiperparámetros y se entrena el modelo utilizando el conjunto de datos de entrenamiento.

5. Fase V. Evaluation.

En esta fase se muestran los resultados obtenidos en la fase V, la primera parte relacionada con el estudio estadístico de los diferentes Datasets, utilizando la regresión lineal como herramienta principal de comparación. La segunda parte está relacionada con los modelos de aprendizaje supervisado aplicados al Dataset llamado “europa_promedio_goles_fifa”, el cual contiene datos sobre promedio de goles y puntos obtenidos antes y después del cambio de entrenador en un equipo de las mejores ligas europeas masculinas.

- ***Parte 1: Resultado de modelos de regresión lineal***

A continuación, se pueden observar las gráficas de las variables utilizadas en la regresión lineal junto con los resultados de cada modelo de regresión lineal

- **Resultados ligas europeas masculinas**
  - Primer modelo (Ligas europeas masculinas)

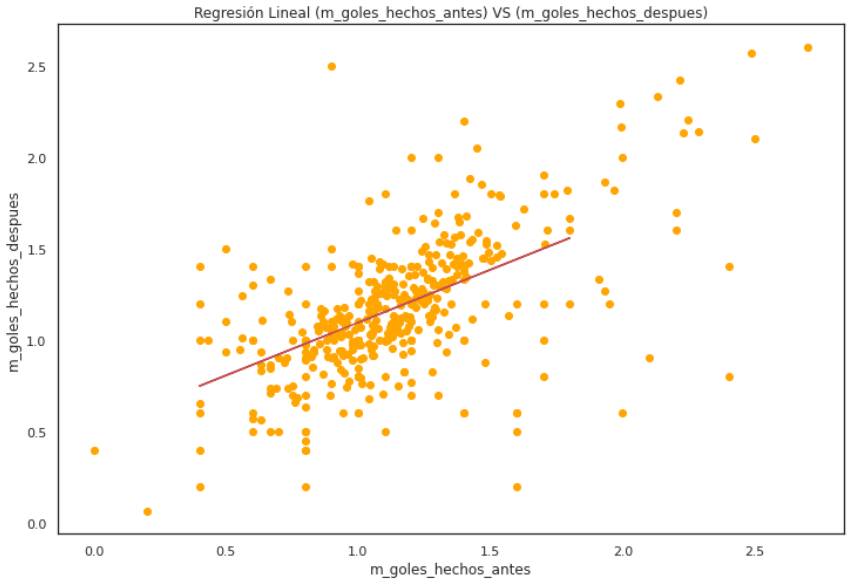

Ilustración 3. Regresión Lineal (m_goles_hechos_antes) VS (m_goles_hechos_despues) ligas europeas masculinas. Fuente: Elaboración propia.

Modelo 1	Europa Ligas masculinas
Variable independiente	m_goles_hechos_antes
Variable dependiente	m_goles_hechos_despues
Ejemplos para entrenamiento	304
Ejemplos para test	76
Coeficiente 1	0.578301
Intersección	0.51694431
Error cuadrático medio	0.06
Error absoluto medio	0.19
R2	0.16

Tabla 13. Resultados del modelo de regresión lineal1 Ligas europeas masculinas. Elaboración propia.

- Segundo modelo (Ligas europeas masculinas)

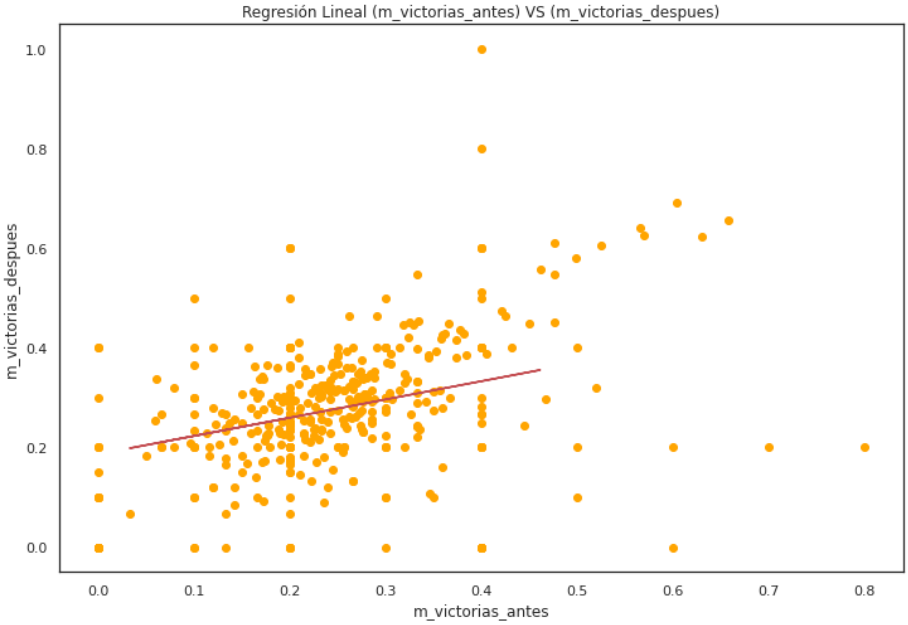

Ilustración 4. Regresión Lineal (m_victorias_antes) VS (m_victorias_despues) ligas europeas masculinas. Fuente: Elaboración propia.

Modelo 2	Europa Ligas masculinas
Variable independiente	m_victorias_antes
Variable dependiente	m_victorias_despues
Ejemplos para entrenamiento	291
Ejemplos para test	73
Coeficiente 1	0.366893
Intersección	0.18628805
Error cuadrático medio	0.01
Error absoluto medio	0.07
R2	0.02

Tabla 14. Resultados del modelo de regresión lineal 2 Ligas europeas masculinas. Elaboración propia.

- Tercer modelo (Ligas europeas masculinas)

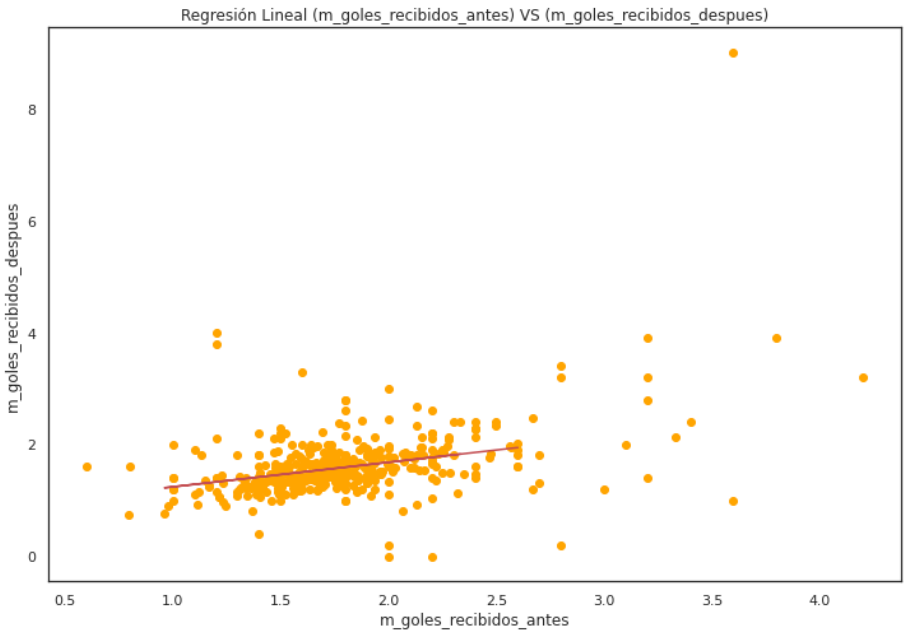

Ilustración 5. Regresión Lineal (m_goles_recibidos_antes) VS (m_goles_recibidos_despues) ligas europeas masculinas. Fuente: Elaboración propia.

Modelo 3	Europa Ligas masculinas
Variable independiente	m_goles_recibidos_antes
Variable dependiente	m_goles_recibidos_despues
Ejemplos para entrenamiento	300
Ejemplos para test	75
Coeficiente 1	0.438593
Intersección	0.80121463
Error cuadrático medio	0.07
Error absoluto medio	0.21
R2	0.24

Tabla 15. Resultados del modelo de regresión lineal 3 Ligas europeas masculinas. Elaboración propia.

- Cuarto modelo (Ligas europeas masculinas)

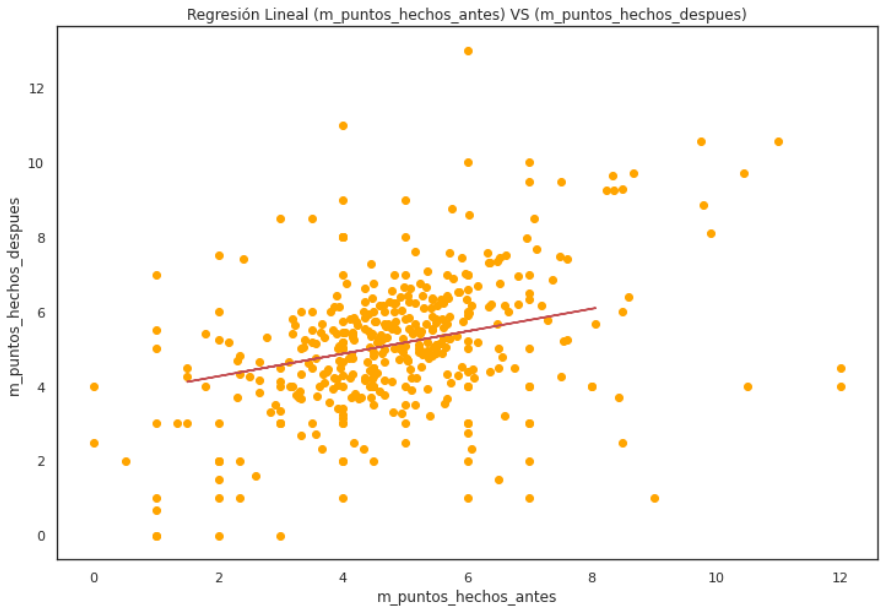

Ilustración 6. Regresión Lineal (m_puntos_hechos_antes) VS (m_puntos_hechos_antes) ligas europeas masculinas. Fuente: Elaboración propia.

Modelo 4	Europa Ligas masculinas
Variable independiente	m_puntos_hechos_antes
Variable dependiente	m_puntos_hechos_despues
Ejemplos para entrenamiento	298
Ejemplos para test	75
Coeficiente 1	0.301747
Intersección	3.66170996
Error cuadrático medio	2.19
Error absoluto medio	1.13
R2	0.08

Tabla 16. Resultados del modelo de regresión lineal 4 Ligas europeas masculinas. Elaboración propia.

- **Ligas Latinoamericanas masculinas**
  - Primer modelo (Ligas Latinoamericanas masculinas)

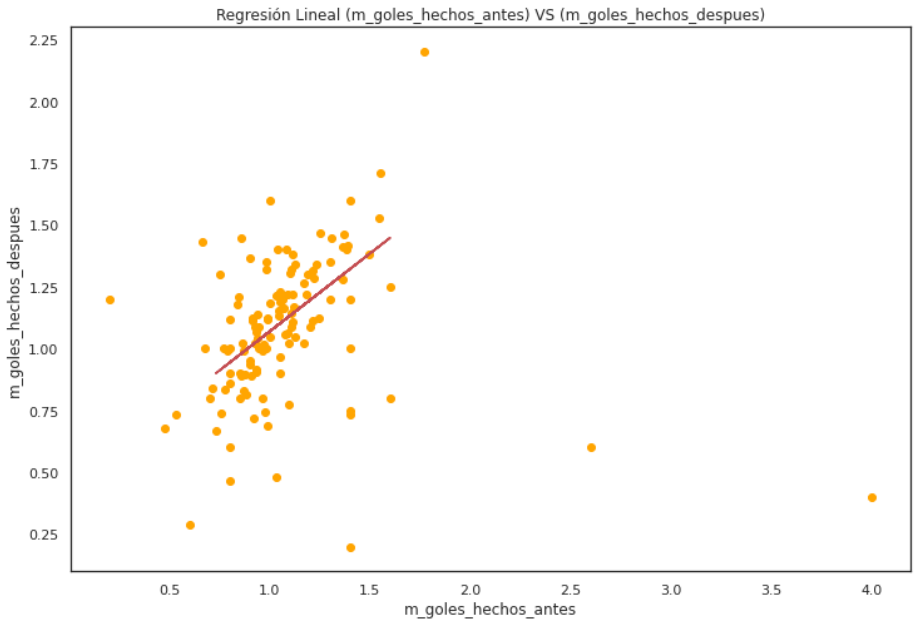

Ilustración 7. Regresión Lineal (m_goles_hechos_antes) VS (m_goles_hechos_despues) ligas latinoamericanas. Fuente: Elaboración propia.

Modelo 1	Latinoamérica Ligas masculinas
Variable independiente	m_goles_hechos_antes
Variable dependiente	m_goles_hechos_despues
Ejemplos para entrenamiento	93
Ejemplos para test	24
Coeficiente 1	0.63101
Intersección	0.43823681
Error cuadrático medio	0.05
Error absoluto medio	0.16
R2	0.01

Tabla 17. Resultados del modelo de regresión lineal 1 Ligas latinoamericanas masculinas. Elaboración propia.

- Segundo modelo (Ligas Latinoamericanas masculinas)

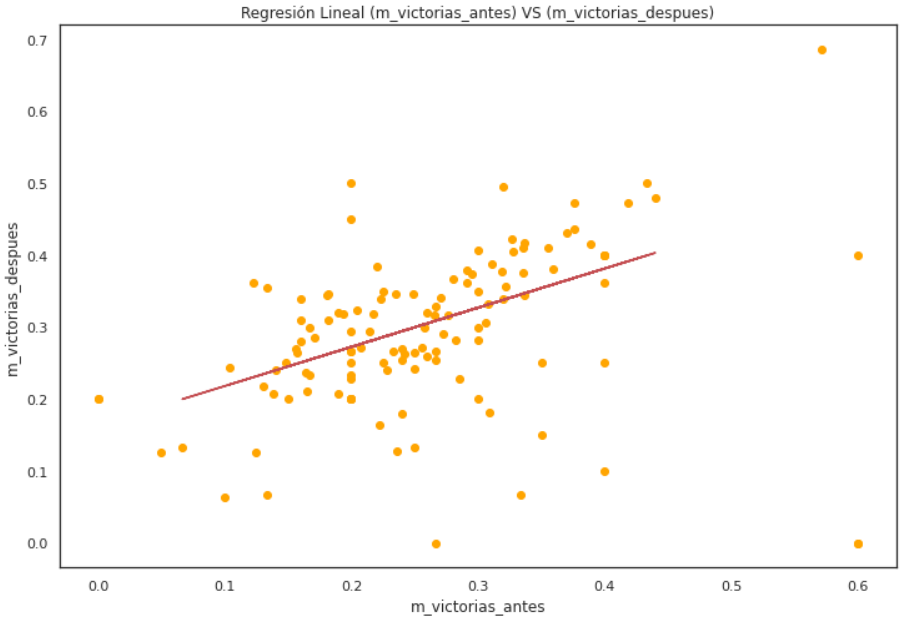

Ilustración 8. Regresión Lineal (m_victorias_antes) VS (m_victorias_despues) ligas latinoamericanas. Fuente: Elaboración propia.

Modelo 2	Latinoamérica Ligas masculinas
Variable independiente	m_victorias_antes
Variable dependiente	m_victorias_despues
Ejemplos para entrenamiento	92
Ejemplos para test	24
Coeficiente 1	0.545025
Intersección	0.16337775
Error cuadrático medio	0.006
Error absoluto medio	0.06
R2	0.31

Tabla 18. Resultados del modelo de regresión lineal 2 Ligas latinoamericanas masculinas. Elaboración propia.

- Tercer modelo (Ligas Latinoamericanas masculinas)

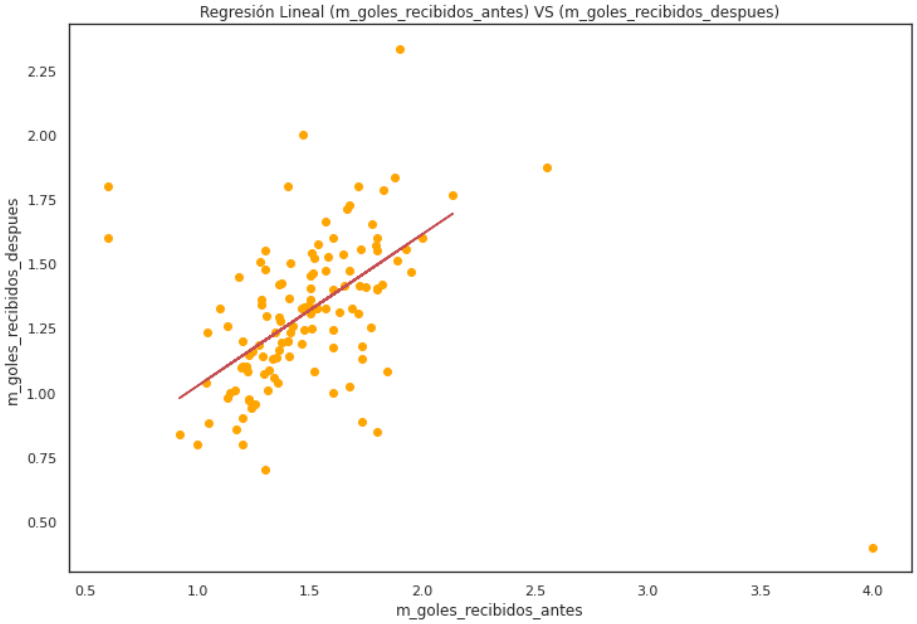

Ilustración 9. Regresión Lineal (m_goles_recibidos_antes) VS (m_goles_recibidos_despues) ligas latinoamericanas. Fuente: Elaboración propia.

Modelo 3	Latinoamérica Ligas masculinas
Variable independiente	m_goles_recibidos_antes
Variable dependiente	m_goles_recibidos_despues
Ejemplos para entrenamiento	94
Ejemplos para test	24
Coeficiente 1	0.589992
Intersección	0.43521369
Error cuadrático medio	0.05
Error absoluto medio	0.16
R2	0.31

Tabla 19. Resultados del modelo de regresión lineal 3 Ligas latinoamericanas masculinas. Elaboración propia.

- Cuarto modelo (Ligas Latinoamericanas masculinas)

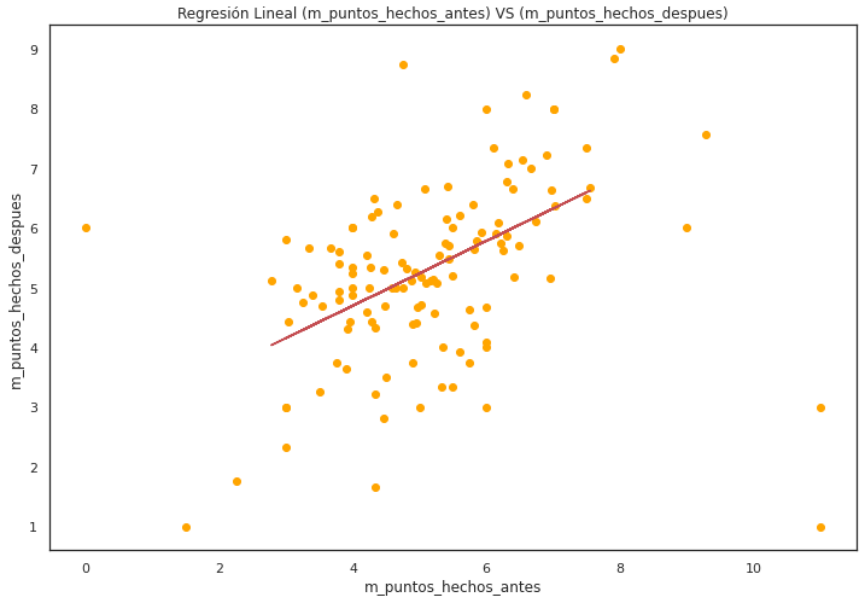

Ilustración 10. Regresión Lineal (m_puntos_hechos_antes) VS (m_goles_hechos_despues) ligas latinoamericanas. Fuente: Elaboración propia.

Modelo 4	Latinoamérica Ligas masculinas
Variable independiente	m_puntos_hechos_antes
Variable dependiente	m_puntos_hechos_despues
Ejemplos para entrenamiento	89
Ejemplos para test	23
Coeficiente 1	0.542455
Intersección	2.52860949
Error cuadrático medio	0.97
Error absoluto medio	0.78
R2	0.43

Tabla 20. Resultados del modelo de regresión lineal 4 Ligas latinoamericanas masculinas. Elaboración propia.

- **Ligas femeninas**
  - Primer modelo (Ligas femeninas)

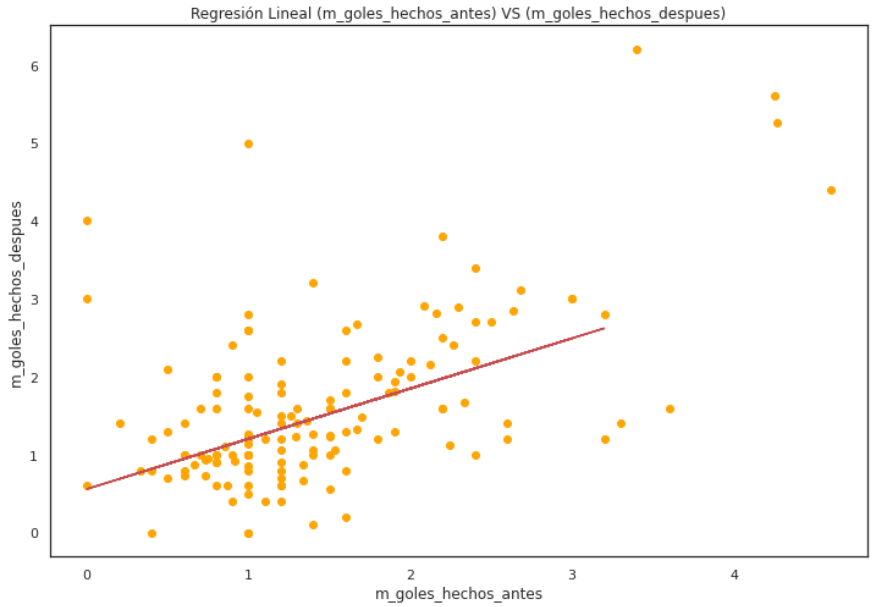

Ilustración 11. Regresión Lineal (m_goles_hechos_antes) VS (m_goles_hechos_despues) ligas femeninas. Fuente: Elaboración propia.

Modelo 1	Europa Ligas femeninas
Variable independiente	m_goles_hechos_antes
Variable dependiente	m_goles_hechos_despues
Ejemplos para entrenamiento	108
Ejemplos para test	28
Coeficiente 1	0.644828
Intersección	0.55807469
Error cuadrático medio	0.47
Error absoluto medio	0.52
R2	0.008

Tabla 21. Resultados del modelo de regresión lineal 1 Ligas europeas femeninas. Elaboración propia.

- Segundo modelo (Ligas femeninas)

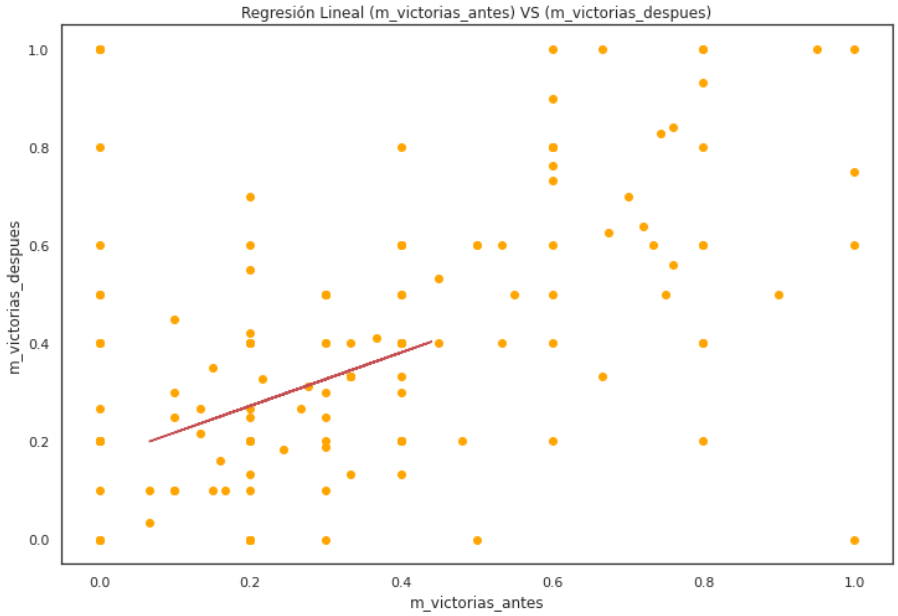

Ilustración 12. Regresión Lineal (m_victorias_antes) VS (m_victorias_despues) ligas femeninas. Fuente: Elaboración propia.

Modelo 2	Europa Ligas femeninas
Variable independiente	m_victorias_antes
Variable dependiente	m_victorias_despues
Ejemplos para entrenamiento	92
Ejemplos para test	24
Coeficiente 1	0.545025
Intersección	0.16337775
Error cuadrático medio	0.006
Error absoluto medio	0.06
R2	0.31

Tabla 22. Resultados del modelo de regresión lineal 2 Ligas europeas femeninas. Elaboración propia.

- Tercer modelo (Ligas femeninas)

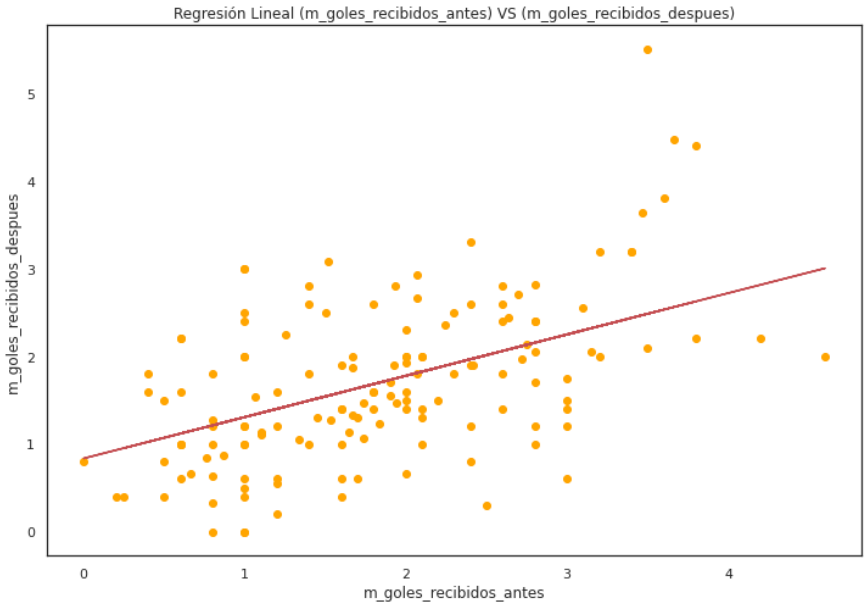

Ilustración 13. Regresión Lineal (m_goles_recibidos_antes) VS (m_goles_recibidos_despues) ligas femeninas. Fuente: Elaboración propia.

Modelo 3	Europa Ligas femeninas
Variable independiente	m_goles_recibidos_antes
Variable dependiente	m_goles_recibidos_despues
Ejemplos para entrenamiento	113
Ejemplos para test	29
Coeficiente 1	0.471353
Intersección	0.83473654
Error cuadrático medio	0.33
Error absoluto medio	0.46
R2	-0.30

Tabla 23. Resultados del modelo de regresión lineal 3 Ligas europeas femeninas. Elaboración propia.

- Cuarto modelo (Ligas femeninas)

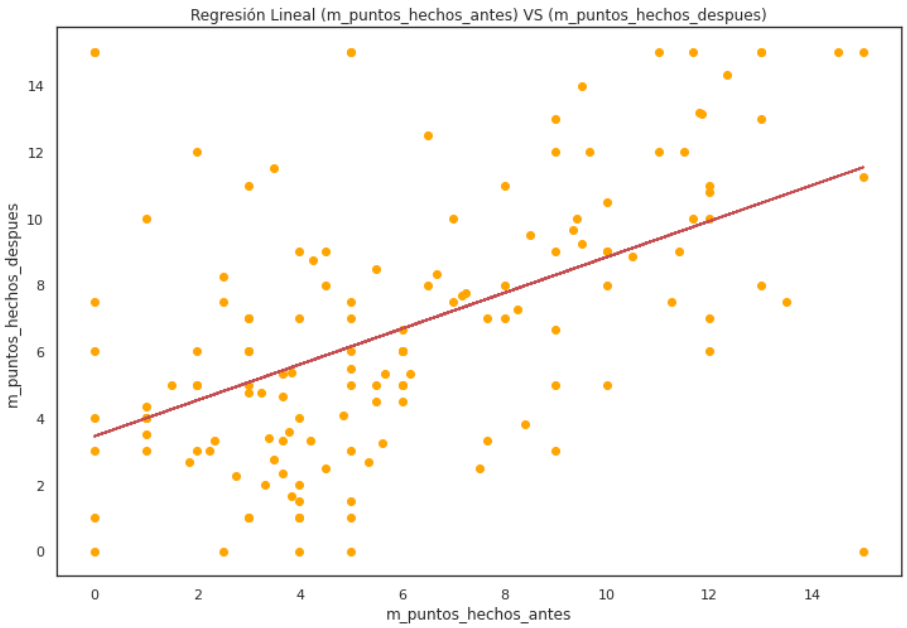

Ilustración 14. Regresión Lineal (m_puntos_hechos_antes) VS (m_puntos_hechos_despues) ligas femeninas. Fuente: Elaboración propia.

Modelo 4	Europa Ligas femeninas
Variable independiente	m_puntos_hechos_antes
Variable dependiente	m_puntos_hechos_despues
Ejemplos para entrenamiento	116
Ejemplos para test	29
Coeficiente 1	0.539062
Intersección	3.45075506
Error cuadrático medio	9.35
Error absoluto medio	2.51
R2	0.33

Tabla 24. Resultados del modelo de regresión lineal 4 Ligas europeas femeninas. Elaboración propia.

- **Discusión de modelos de regresión lineal**

El análisis estadístico revela que en general no hay diferencia significativa entre las ligas masculinas y femeninas, exacto por la particularidad de que las ligas femeninas poseen medias más elevadas, sin embargo, el incremento de beneficios que reciben posterior al cambio de entrenador es similar al de las ligas masculinas. También es importante aclarar que todas las variables poseen una alta dispersión, razón por la cual los modelos de regresión lineal que se aplicaron sobre las variables afines, presentan bajos coeficientes de determinación. Sin embargo, la finalidad de la regresión lineal en este caso es detectar la tendencia de los datos, por ejemplo, los modelos realizados para las variables referentes a la media de goles hechos antes y después del cambio de entrenador determinaron que los equipos que anotaban más de 3 goles por partido en promedio tienden a no beneficiarse demasiado del cambio de entrenador. Mientras que los equipos con un predio de goles inferior a 3 puede aumentar este indicador posterior al cambio de entrenador en mayor medida. Lo anterior da la siguiente interpretación, si un equipo se encuentra en buen rendimiento en el momento del cambio de entrenador, no aumentará en mayor medida dicho rendimiento con el nuevo entrenador, mientras que, si el equipo anota pocos goles o ningún gol en promedio por partido, entonces tendrá un incremento mayor en sus resultados posterior al cambio de entrenador. Las demás variables afines a estudio presentaron el mismo patrón, pues según la regresión existe un punto en el que algunos equipos no se benefician del efecto del cambio de entrenador, mayormente los equipos que poseen un buen rendimiento en esa variable en particular, también cabe destacar que los modelos de regresión aplicados a un par de variables en los tres tipos de ligas presentaron valores similares, variando solo en la intercepción de la regresión lineal.

Las gráficas de las variables afines demuestran una tendencia de los datos a la linealidad, aunque existe una alta variabilidad en todas las características, la regresión lineal es útil para observar el comportamiento de los datos. Se tenía un breve pronostico del posible comportamiento de los datos en la revisión de las medias de las variables donde se pudo observar cómo aumentaba ligeramente cada variable luego del cambio de entrenador, los coeficientes de la regresión lineal demuestran un comportamiento similar entre las mismas características de los tres Datasets, en primer lugar las variables “m_goles_hechos_antes” (Variable Independiente) y “m_goles_hechos_despues” (Variable dependiente) poseen los coeficientes 0.57, 0.63, 0.64 y la intersección 0.51, 0.43 0.55 para las ligas Europeas masculinas, ligas latinoamericanas y ligas femeninas respectivamente.

Los coeficientes de los tres Dataframes son similares sobre todo en las ligas latinoamericanas y las ligas femeninas. Las intersecciones de los tres Dataframes también presentan similitud en los valores, las ligas europeas masculinas tienen una magnitud levemente más pequeña en el coeficiente lo que indica que la línea de la regresión lineal escala en menor medida que las regresiones realizadas para las ligas y latinoamericanas masculinas y las ligas femeninas, sin embargo, todos los coeficientes poseen magnitudes superiores a 0.5. En las gráficas se puede observar que los tres modelos se comportan de forma similar, indicando que los equipos que anotan pocos

goles antes del cambio de entrenador pueden aumentar su rendimiento levemente pero las condiciones cambian para los equipos que anotan más de 3 goles en promedio por partido ya que se disminuye el promedio de goles, el coeficiente de determinación indica que el modelo que más se ajustó a los datos fue el modelo para las ligas europeas masculinas con un valor de 0.16, sin embargo el modelo que posee el menor MAE es el modelo de las ligas femeninas.

El modelo 2 de regresión de cada Dataframe fue construido con las variables "m_victorias_antes" (Variable independiente), y "m_victorias_despues" (Variable dependiente), donde el coeficiente para cada modelo fue de 0.36, 0.54, 054 y la intersección para cada modelo fue de 0.18, 0.16 y 0.16 para las ligas europeas masculinas, ligas latinoamericanas y ligas femeninas respectivamente. Los coeficientes de las ligas latinoamericanas y ligas femeninas son iguales o superiores a 0.5, lo cual indica que los equipos que cambian de entrenador en estas ligas suelen en promedio aumentar o mantener el porcentaje de victorias que tenían en los 5 partidos previos al cambio de entrenador, mientras que las ligas europeas poseen una magnitud de coeficiente de 0.36, aunque hay mucha dispersión en los datos, este valor indica que de hecho el cambio de entrenador puede llegar a ser no beneficioso a corto plazo, pues se reduce el porcentaje de victorias. Los modelos generados de las ligas latinoamericanas y femeninas obtuvieron el mismo coeficiente de determinación el cual fue de 0.31, mientras que para las ligas europeas fue de tan solo 0.02, como las ligas latinoamericanas y femeninas también poseen en este caso el mismo MAE y MSE, el cual fue de 0.06 y 0.006 respectivamente.

Las variables para el modelo 3 de cada Dataframe fueron "m_goles_recibidos_antes" (Varaible independiente) y "m_goles_recibidos_despues" (variable independiente), los coeficientes para cada modelo fueron 0.43, 0.58, 0.47 y las intersecciones para cada modelo fueron 0.80, 0.43, 0.83 respectivamente.

En la sección de análisis de medias se observó como las variables referentes a los goles que un club deportivo recibe cuando cambia de entrenador a corto plazo se presentaban una disminución con respecto a los goles recibidos antes del cambio, los coeficientes demuestran concordancia con este hecho dado que las intercesiones de los modelos pertenecientes a las ligas europeas masculinas y ligas femeninas están cerca de una unidad completa, pero con coeficientes inferiores a 0.5, lo que indica que se reduce la cantidad de goles recibidos en promedio cuando se realiza el cambio de entrenador.

Para las ligas latinoamericanas sucede de la misma manera aunque la intercepción posee una menor magnitud, el coeficiente de la variable independiente es mayor, en el análisis de medias y medianas se observó como Latinoamérica presentaba también la menor media y mediana en comparación con las ligas europeas masculinas y ligas femeninas para las variables "goles_recibidos_antes_cambio_entrenador" y "goles recibidos_despues_cambio_entrenador" (1.43 y 1.26 respectivamente), aunque en promedio sean menores los valores, el efecto es similar al de las ligas europeas y femeninas, el modelo de regresión para las ligas femeninas posee un coeficiente de determinación de -0.30, esta medida puede crear confusión ya que se podría llegar a pensar que la correlación entre las dos variables es negativa, lo cual no es el caso, sucede que los datos de entrenamiento los cuales son el ultimo 20% de los datos

disponibles para el entrenamiento se encuentran inversos a lo que dicta el modelo de regresión, dando como resultado el valor negativo, en este caso es factible guiarse por las dos métricas de error restantes, el modelo con el MAE y el MSE más bajo fue el modelo de las ligas latinoamericanas con 0.16, y 0.05 respectivamente.

Para el modelo 4 el cual es el modelo final aplicado a los tres Dataframes utilizando las variables "m_puntos_hechos_antes" (Variable independiente) y "m_puntos_hechos_despues" (Variable dependiente), donde los coeficientes de los modelos tienen el valor de 0.30, 0.54, 0.53 y las intersecciones tienen el valor de 3.66, 2.52, 3.45, valores obtenidos para los modelos creador a partir de los Dataframes de las ligas europeas masculinas, ligas latinoamericanas y ligas femeninas respectivamente.

La intersección para el modelo de las ligas europeas es el más alto aunque posee el coeficiente de variable independiente más bajo, la regresión del modelo de las ligas europeas posee una pendiente con menor magnitud que los modelos de las ligas latinoamericanas y femeninas, por lo tanto tiene una pendiente con menor magnitud, según el modelo, los equipos que generan más de 4.5 puntos en promedio tenderán a aumentar dichos puntos con el cambio de entrenador mientras que los equipos que generen más de 4.5 puntos tenderán a mantener o reducir dicho valor después del cambio de entrenador.

El modelo para las ligas femeninas tiene una intersección similar a al modelo de las ligas europeas, pero con un coeficiente de variable independiente de 0.53, lo cual genera una línea de regresión con una pendiente casi de 45 grados, indicando que los equipos que generen menos de 7.5 puntos aproximadamente antes de realizar el cambio de entrenador, tenderán a incrementar la generación de puntos a corto plazo, luego de los 7.5 puntos comenzaran a generar igual o menos puntos de los que generaban previo al cambio pero en menor medida que las ligas europeas masculinas, finalmente el modelo para las ligas latinoamericanas presenta la intersección más baja pero el coeficiente de variable más alto, se genera un efecto similar al de los modelos de las ligas europeas masculinas y ligas femeninas, en este caso los equipos que generen menos de 5 puntos en promedio aumentaran su generación de puntos luego del cambio de entrenador, el modelo con el índice de determinación más elevado es el de las ligas latinoamericanas con 0.43, debido a que en este grupo de muestras, los datos parecen agruparse dejando pocos Outliers a los alrededores, también obtuvo el MAE y MSE más bajo con 0.97 y 0.78 respectivamente.

- ***Parte 2: Resultados de aplicación de modelos de aprendizaje supervisados***

- **Resultados Modelos de aprendizaje aplicados a la variable "puntos_hechos_despues_de_cambio_entrenador_1_5"**

- RandomForests para predicción de puntos obtenidos después del cambio de entrenador (Resultado)

Luego de aplicar el modelo RandomForests al conjunto de datos de entrenamiento buscando predecir la variable “puntos_hechos_despues_de_cambio_entrendor_1_5” se obtiene la siguiente gráfica:

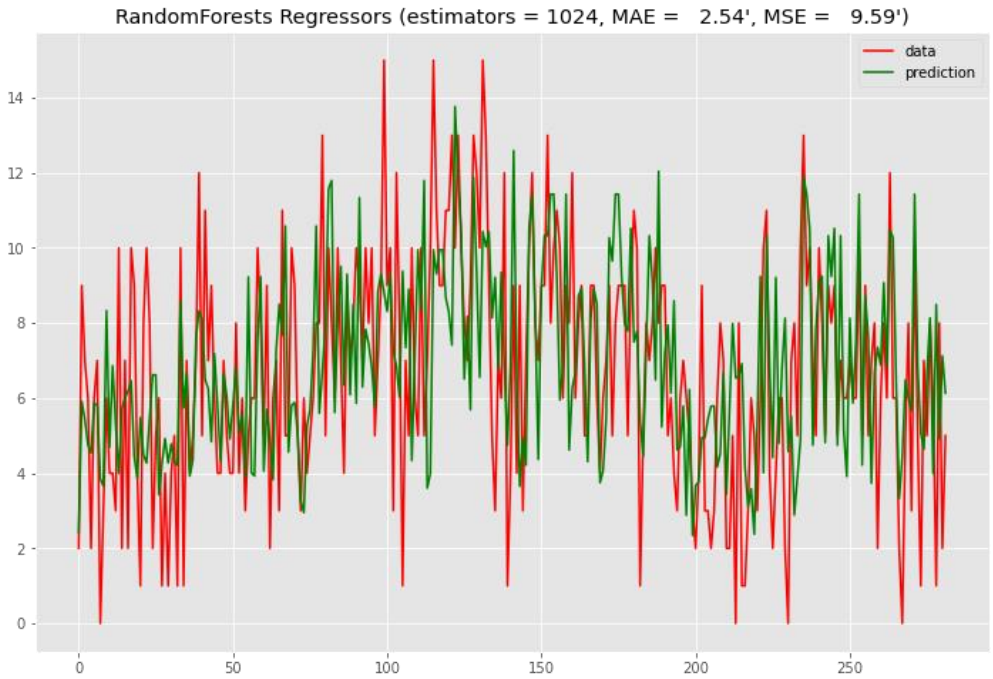

Ilustración 15 .Comparación de predicciones generadas por el modelo RandomForests y valores reales (variable objetivo: puntos_hechos_despues_cambio_entrenador). Elaboración propia.

Modelo	RandomForests
Variable objetivo	puntos_hechos_despues_de_cambio_entrenador_1_5
MAE	2.54
MSE	9.59

Tabla 25. Resultado de métricas del modelo RandomForest (variable: puntos_hechos_despues_cambio_entrenador). Fuente: Elaboración propia.

La grafica presenta un MAE de 2.54, los valores de color rojo representan los datos reales, mientras que los verdes representan las predicciones realizadas por el modelo.

En la gráfica 15 se puede observar que el modelo trata de ajustarse para lograr obtener los puntos que un equipo obtendrá en los siguientes 5 encuentros si cambia de entrenador basado en la jornada que se encuentre y los puntos obtenidos en los 5 partidos previos al cambio de entrenador, el modelo se ajusta a los valores lo mejor que puede excepto en los picos donde no logra predecir con exactitud los valores reales, esta diferencia entre los picos altos y la predicción genera el MAE de 2,54.

Con el atributo “feature_importances_” se puede observar la relevancia de las características que afectan el modelo.

Variable	Relevancia
jornada	0.216637
puntos_actuales	0.264847
puntos_hechos_antes_de_cambio_entrenador_1_5	0.228409
ataque	0.290107

Tabla 26. Relevancia de las características del modelo RandomForest (variable objetivo: puntos_hechos_despues_cambio_entrenador). Elaboración propia.

La característica “ataque”, es la más influyente de las variables sin embargo no esta tan alejada de la variable “puntos actuales”, la característica con menos influencia en el modelo es la variable “jornada”, aunque en esta ocasión no está muy alejada de las demás características, todas están niveladas casi a las mismas cifras.

- DecisionTree para predicción de puntos obtenidos después de cambio de entrenador (Resultado)

Para el modelo de árbol de decisión prediciendo la variable “puntos_hechos_despues_de_cambio_entrendor_1_5” se obtiene la siguiente gráfica:

Ilustración 16. Comparación de predicciones generadas por el modelo DecisionTree y valores reales (variable objetivo: puntos_hechos_despues_cambio_entrenador). Fuente: Elaboración propia.

Modelo	DecisionTree
Variable objetivo	puntos_hechos_despues_de_cambio_entrenador_1_5
MAE	2.40
MSE	8.86

Tabla 27. Resultado de métricas del modelo DecisionTree (variable objetivo: puntos_hechos_despues_cambio_entrenador). Elaboración propia.

El grafico 16 es un gráfico similar generado por el modelo RandomForests, donde se obtiene un MAE de 2.40, el árbol de decisión presenta un menor error absoluto medio, el modelo de árbol de decisión genera predicciones que en su mayoría no descienden de 4 puntos, tiende a estar sobre la media de puntos hechos antes del cambio de entrenador de todo el conjunto de datos la cual es de 5.35.

Variable	Relevancia
jornada	0.235132
puntos_actuales	0.525881
puntos_hechos_antes_de_cambio_entrenador_1_5	0.052313
ataque	0.186674

Tabla 28. Relevancia de las características del modelo DecisionTree (variable objetivo: puntos_hechos_despues_cambio_entrenador). Elaboración propia.

Para el modelo de árbol de decisión, la característica “puntos_actuales” es la característica que más influye en el modelo, seguida de la característica “jornada”, la característica menos importante es la referente a el número de puntos hechos antes del cambio de entrenador.

- AdaBoost para predicción de puntos obtenidos después de cambio de entrenador (Resultado)

Luego de aplicar el modelo AdaBoost sobre los resultados buscando predecir la variable “puntos_hechos_despues_de_cambio_entrendor_1_5”, se obtiene la siguiente gráfica:

Ilustración 17. Comparación de predicciones generadas por el modelo AdaBoost y valores reales (variable: puntos_hechos_despues_cambio_entrenador). Fuente: Elaboración propia.

Modelo	AdaBoost
Variable objetivo	puntos_hechos_despues_de_cambio_entrenador_1_5
MAE	2.40
MSE	8.67

Tabla 29. Resultado de métricas del modelo AdaBoost (variable: puntos_hechos_despues_cambio_entrenador). Elaboración propia.

La técnica de AdaBoost aplicada al conjunto de datos no presenta picos demasiado altos, inclusive las predicciones generadas no sobrepasan los 10 puntos en su mayoría.

El resultado de las predicciones con los datos reales refleja un MAE de 2.40.

Variable	Relevancia
jornada	0.268230
puntos_actuales	0.359110
puntos_hechos_antes_de_cambio_entrenador_1_5	0.133747
Ataque	0.238913

Tabla 30. Relevancia de las características del modelo Adaboost (variable: puntos_hechos_despues_cambio_entrenador). Elaboración propia.

En esta ocasión la variable más importante fue la variable “puntos_actuales”, seguida de la variable “jornada”, para este modelo la variable “porcentaje _victorias_antes_de_cambio_entrenador_1_5” vuelve a ser la menos relevante.

- **GradientBoosting para predicción de puntos después de cambio de entrenador (Resultado).**

Para el modelo de GradientBoosting obtenemos la siguiente grafica comparando las predicciones con los datos reales de la variable “puntos_hechos_despues_de_cambio_entrendor_1_5”:

Ilustración 18. Comparación de predicciones generadas por el modelo GradientBoosting y valores reales (variable objetivo: puntos_hechos_despues_cambio_entrenador). Elaboración propia.

Modelo	GradientBoosting
Variable objetivo	puntos_hechos_despues_de_cambio_entrenador_1_5
MAE	2.40
MSE	8.76

Tabla 31. Resultado de métricas del modelo Gradientboosting (variable: puntos_hechos_despues_cambio_entrenador). Fuente: Elaboración propia.

El GradientBoosting es el último modelo para predecir la variable “puntos_hechos_despues_de_cambio_entrendor_1_5”,

El MAE para generado por la comparación de las predicciones y los datos reales es de 2.54.

Variable	Relevancia
jornada	0.250907
puntos_actuales	0.298988
puntos_hechos_antes_de_cambio_entrenador_1_5	0.154040
ataque	0.296066

Tabla 32. Relevancia de las características del modelo GradientBoosting (variable objetivo: puntos_hechos_despues_cambio_entrenador). Elaboración propia.

La característica más importante para el modelo de GradientBoosting prediciendo la variable "puntos_hechos_despues_de_cambio_entrenador" es "puntos_actuales" con el 29.8% de la importancia, y la segunda variable más importante fue "ataque" con 29.6%, las dos variables restantes no presentan mayor relevancia para este modelo, ya que juntas representan apenas el 13%.

- **Resultado Modelos de aprendizaje aplicados a la variable "promedio_goles_hechos_despues_de_cambio_entrenador_1_5"**

    - RandomForests para predicción de goles hechos después del cambio de entrenador (Resultado)

Posterior a la aplicación del modelo de RandomForests a el conjunto de datos, esta vez para predecir la variable "promedio_de_goles_hechos_despues_de_cambio_entrendor_1_5", se obtiene la siguiente gráfica:

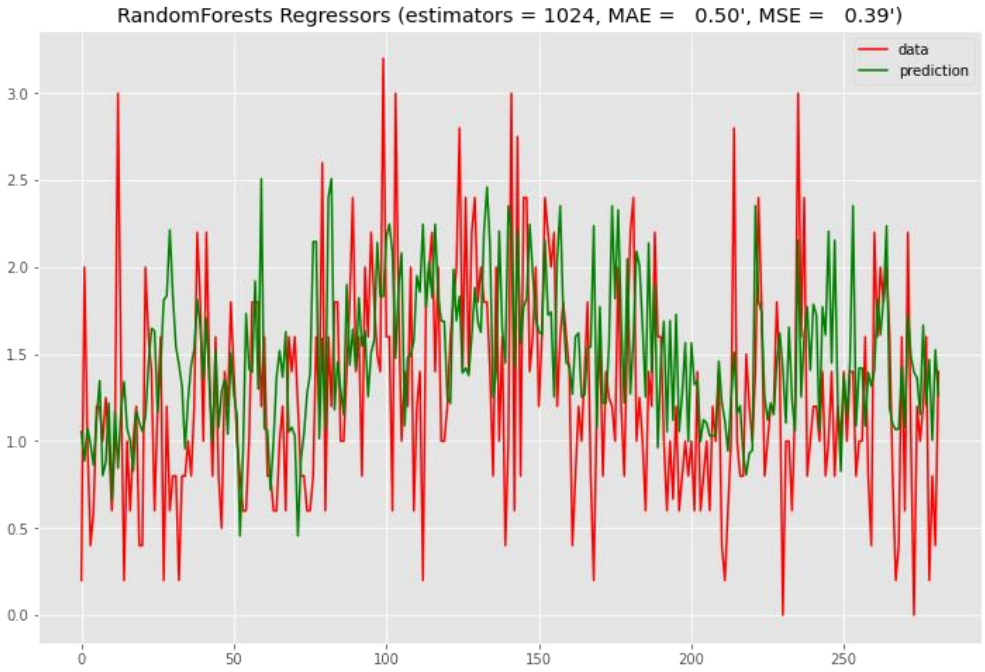

Ilustración 19. Comparación de predicciones generadas por el modelo RandomForests y valores reales (variable objetivo: promedio_goles_hechos_despues_cambio_entrenador). Fuente: Elaboración propia.

Modelo	RandomForests
Variable objetivo	promedio_goles_hechos_despues_de_cambio_entrenador_1_5
MAE	0.5
MSE	8.86

Tabla 33. Resultado de métricas del modelo RandomForest (variable: promedio_goles_hechos_despues_de_cambio_entrenador_1_5). Fuente: Elaboración propia.

En la gráfica 19 presenta un valor de MAE de 0.50 aplicando el modelo RandomForests comparando las predicciones con los datos reales.

Variable	Relevancia
jornada	0.209244
puntos_actuales	0.249507
promedio_goles_hechos_antes_de_cambio_entrenador_1_5	0.225710
ataque	0.315539

Tabla 34. Relevancia de las características del modelo RandomForest (variable objetivo: promedio_goles_hechos_despues_de_cambio_entrenador_1_5). Fuente: Elaboración propia.

Utilizando el RandomForests para este conjunto de datos, se puede observar que la relevancia de las características esta nivelada, la característica con más importancia es la de “ataque” con 27% aunque su importancia es cercana a la de las demás caracteristicas.

- DecisionTree para Predicción de goles hechos después de cambio de entrenador.

Se aplica el modelo de árbol de decisión al conjunto de datos de entrenamiento y se obtienen las predicciones para la variable "promedio_de_goles_hechos_despues_de_cambio_entrendor_1_5", donde se obtiene la siguiente gráfica:

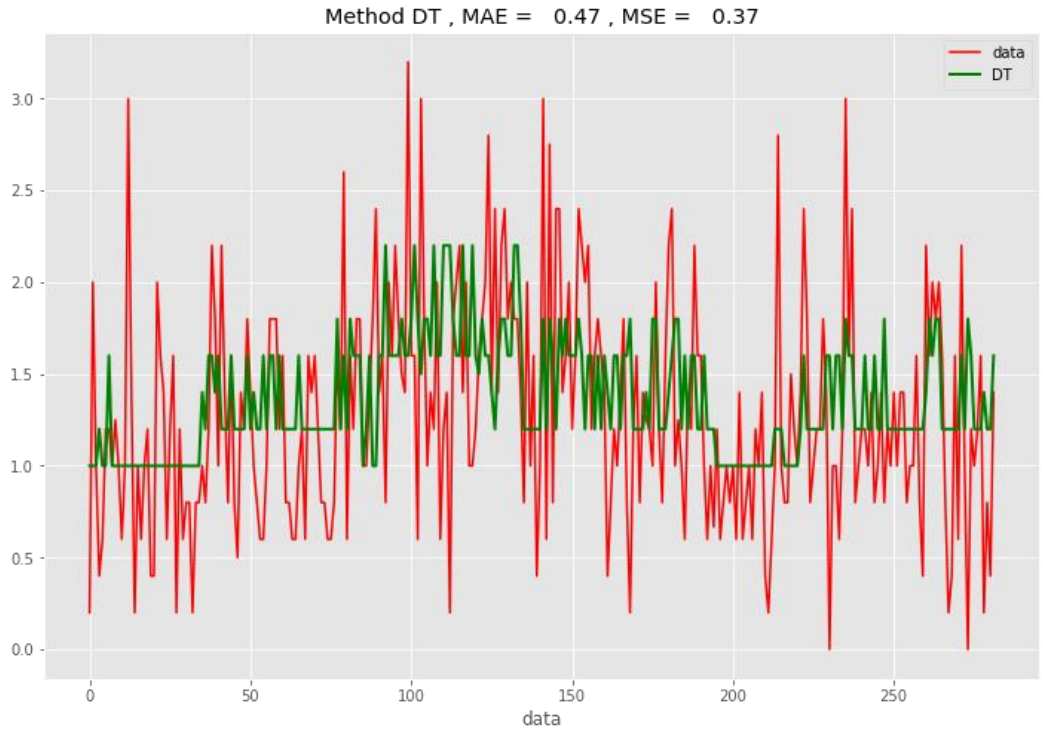

Ilustración 20. Comparación de predicciones generadas por el modelo DecisionTree y valores reales (variable objetivo: promedio_goles_hechos_despues_cambio_entrenador). Fuente: Elaboración propia.

Modelo	DesicionTree
Variable objetivo	Promedio_goles_hechos_despues_de_cambio_entrenador_1_5
MAE	0.47
MSE	0.37

Tabla 35. Resultado de métricas del modelo DecisionTree (variable: promedio_goles_hechos_despues_de_cambio_entrenador_1_5). Elaboración propia.

El modelo finaliza obteniendo un MAE de 0.47, el modelo de árbol de decisión es más sencillo que los conjuntos de modelos, aunque suele obtener buenos resultados también, en este caso obtiene un MAE inferior al obtenido por el modelo de RandomsForests, sin embargo, en la gráfica observamos que las predicciones son más centralizadas, se mantienen centradas en la mayoría sobre el promedio de goles 1 y de este valor no disminuye.

Variable	Relevancia
jornada	0.072275
puntos_actuales	0.174566
promedio_goles_hechos_antes_de_cambio_entrenador_1_5	0.414100
ataque	0.339060

Tabla 36. Relevancia de las características del modelo DecisionTree (variable objetivo: promedio_goles_hechos_despues_de_cambio_entrenador_1_5). Elaboración propia.

En la selección de características notamos que la variable con más impacto es la variable "promedio_goles_hechos_antes_de_cambio_entrenador_1_5", en esta ocasión no se posee un equilibro en la importancia de las variables con en el caso del RandomForests, la variable "jornada" es la variable menos relevante, en varios modelos aplicados para predecir la variable "promedio_de_goles_hechos_despues_de_cambio_entrendor_1_5" , se obtuvo un valor de relevancia alto para la característica "jornada", en esta ocasión es totalmente opuesto.

- **AdaBoost para Predicción de goles hechos después de cambio de entrenador**

Se entrena el modelo con el conjunto de datos y se generan las predicciones para la variable "promedio_de_goles_hechos_despues_de_cambio_entrenador_1_5", se obtiene la siguiente gráfica:

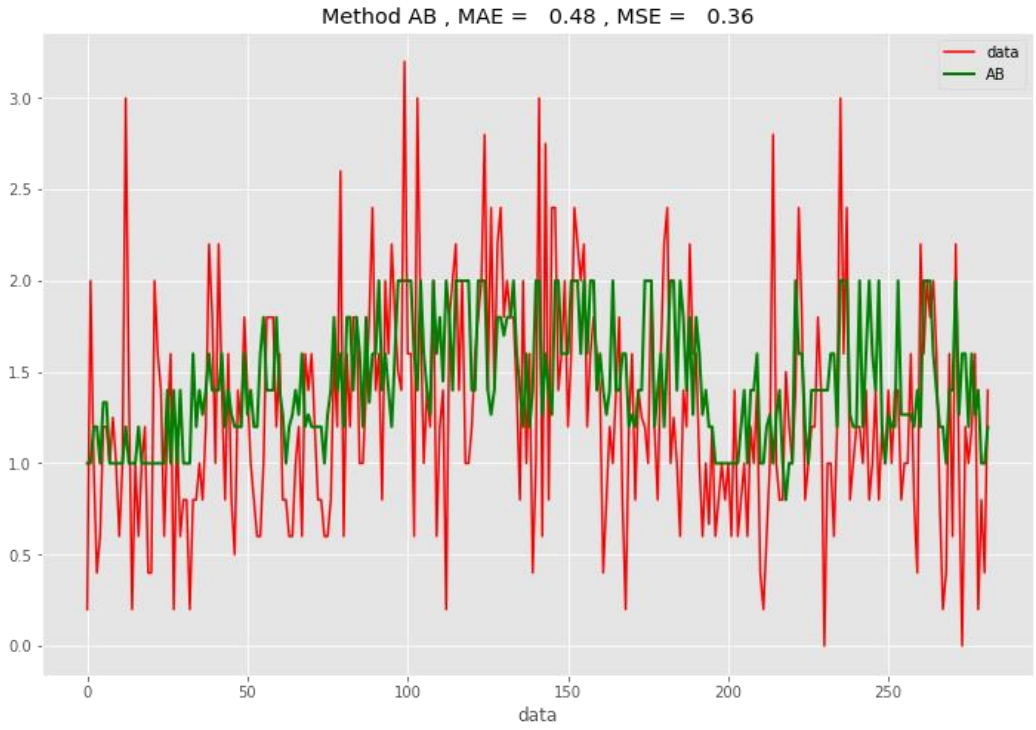

Ilustración 21. Comparación de predicciones generadas por el modelo AdaBoost y valores reales (variable: promedio_goles_hechos_despues_cambio_entrenador). Fuente: Elaboración propia.

Modelo	AdaBoost
Variable objetivo	Promedio_goles_hechos_despues_de_cambio_entrenador_1_5
MAE	0.48
MSE	0.36

Tabla 37. Resultado de métricas del modelo AdaBoost (variable objetivo: promedio_goles_hechos_despues_de_cambio_entrenador_1_5). Elaboración propia.

El resultado de comparar las predicciones con los valores reales genera un MAE de 0.48, este modelo genera una respuesta similar al modelo de árbol de decisión ya que no genera predicciones para el promedio de goles hechos por debajo de 1 gol.

Variable	Relevancia
jornada	0.206894
puntos_actuales	0.336565
promedio_goles_hechos_antes_de_cambio_entrenador_1_5	0.214479
ataque	0.242091

Tabla 38. Relevancia de las características del modelo AdaBoost (variable objetivo: promedio_goles_hechos_despues_de_cambio_entrenador_1_5). Elaboración propia.

La variable "promedio_goles_hechos_antes_de_cambio_entrenador_1_5" posee la mayor relevancia (33% de relevancia), mientras que la variable "jornada" tiene la menor influencia en el modelo (20% de relevancia).

- **GradientBoosting para Predicción de goles hechos después de cambio de entrenador.**

El último modelo que se probó utilizando conjunto de datos de testeo genera la siguiente gráfica teniendo como variable objetivo "promedio_goles_hechos_antes_de_cambio_entrenador_1_5":

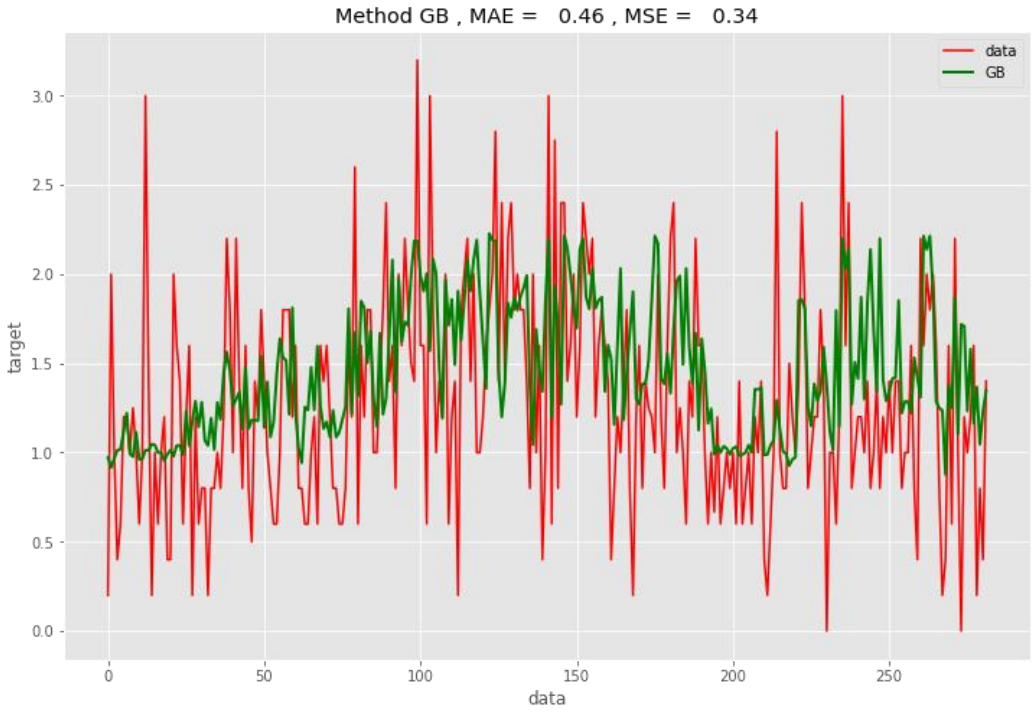

Ilustración 22. Comparación de predicciones generadas por el modelo GradientBoosting y valores reales (variable: promedio_goles_hechos_despues_cambio_entrenador). Fuente: Elaboración propia.

Modelo	GradientBoosting
Variable objetivo	Promedio_goles_hechos_despues_de_cambio_entrenador_1_5
MAE	0.46
MSE	0.34

Tabla 39. Resultado de métricas del modelo Gradientboosting (variable: promedio_goles_hechos_despues_de_cambio_entrenador_1_5). Elaboración propia.

El valor generado por la comparación entre los valores reales y las predicciones es de 0.50, lo que presenta el mejor MAE obtenido para la variable objetivo "promedio_goles_hechos_antes_de_cambio_entrenador_1_5", este modelo comparte algunas similitudes con el modelo de árbol de decisión y el AdaBoost, entre ellas que la mayoría de las predicciones se mantienen entre un intervalo de 1 y 2 goles de promedio en los 5 partidos posteriores al cambio de entrenador, a diferencia del RandomForests que realiza predicciones más elevadas.

Variable	Relevancia
jornada	0.161425
puntos_actuales	0.293120
promedio_goles_hechos_antes_de_cambio_entrenador_1_5	0.285012
ataque	0.260442

Tabla 40. Relevancia de las características del modelo GradientBoosting (variable: promedio_goles_hechos_despues_de_cambio_entrenador_1_5). Elaboración propia.

Por último, la variable “puntos_actuales”, es la variable más alta, con el 29% de la importancia, mientras que la variable “jornada”, es la variable con menos importancia con el 16%, sin embargo, todas las variables demuestran relevancia, lo cual es un buen indicativo.

- **Discusión segunda parte**

Luego de haber construido los modelos de aprendizaje supervisado, entrenarlos y probarlos con los conjuntos de testeo correspondientes para predecir la variable “puntos_hechos_despues_de_cambio_entrenador_1_5”, se obtuvieron en los cuatro modelos, resultados muy cercanos entre ellos, los cuales rondaban el valor MAE de 2.40, excepto para el modelo de RandomForests el cual fue de 2.54. Los resultados indican que los modelos tendrán una diferencia de aproximadamente 2.4 puntos entre el valor real y la predicción generada, lo cual es un número alejado del valor real.

Los modelos tratan predecir los resultados utilizando variables relacionadas con datos anteriores al cambio de entrenador, como el promedio de goles hechos en los últimos 5 partidos anteriores al cambio de entrenador o como el ataque del equipo o la jornada del cambio de entrenador.

La variable más relevante en todos los modelos fue “puntos_actuales”, excepto en el modelo RandomForest donde la variable más relevante fue “ataque”, sin embargo, en este modelo la variable “puntos_actuales” tuvo el segundo lugar con el 26% de relevancia, lo cual indica que la variable “puntos_actuales” sea la variable independiente más determinante en la predicción de la variable dependiente “puntos_hechos_despues_de_cambio_entrenador”, por otro lado, no hay una sola variable que haya quedado en último lugar en los cuatro modelos, pues todas las variables excepto “puntos_actuales” destacaron en algún modelo como la variable menos relevante, sin embargo la variable que menos porcentaje de relevancia tuvo en todos los modelos fue la variable “puntos_hechos_antes_de_cambio_entrenador_1_5”, en el modelo DecisionTree con 5% de relevancia, aunque este hecho ocurrió solo en ese modelo, en los demás modelos tuvo un desempeño a la par con las demás variables rondando el 20% de relevancia.

Para el caso de los modelos que generan las predicciones de la variable “promedio_goles_hechos_despues_cambio_entrenador_1_5”, el mejor resultado lo brindo el modelo GradientBoosting con un MAE de 0.46, la variable con más relevancia de los modelos fue la variable “puntos_actuales”, obteniendo el primer lugar en los modelos AdaBoost y GradientBoosting, mientras que la variable “jornada”, fue la variable con menos relevancia, ya que obtuvo el último puesto de relevancia en todos los modelos, siendo su peor desempeño en el modelo DecisionTree con el 7% de relevancia, sin embargo, en el resto de modelos estuvo nivelado con el resto de variables.

Cómo se analizó en el la fase III, las variables generadas con los datos obtenidos de los encuentros anteriores y posteriores a un cambio de entrenador relacionadas pero presentan una varianza alta, esto resulta en que las variables más importantes en varios modelos sean “jornada” y “puntos_actuales”, que son variables tomadas justo en el momento del cambio de entrenador, la variable “promedio_goles_hechos_despues_de_cambio_entrenador_1_5” tiene alta correlación con la variable “promedio_goles_hechos_antes_de_cambio_entrenador_1_5”, al igual que la variable “puntos_hechos_antes_de_cambio_entrenador_1_5” tiene alta correlación con la variable “promedio_victorias_despues_de_cambio_entrenador_1_5”, pero la variabilidad de los datos hace que sea difícil llegar a predecir las variables objetivo seleccionadas con más exactitud.

En los modelos de aprendizaje supervisado no es posible incluir variables como características para entrenar que tengan en su nombre “después”, debido a que estas variables contienen datos del futuro pues son información obtenida de los partidos posteriores al cambio de entrenador, por ejemplo, no es posible implementar la variable “promedio_victorias_despues_de_cambio_entrenador_1_5” en alguno de los modelos pues se estaría filtrando información lo que concluiría en que los modelo tenderían a tener Overfitting lo que no sería de mucha utilidad, ya que cada modelo siempre requeriría que se le brindase dicha variable que no existiría para un caso real.

Para este caso en particular, la tarea de predecir el número de puntos o goles es laboriosa ya que como se ha observado en este trabajo, no solo conocer los datos anteriores al cambio de entrenador asegura que se vaya a tener una buena predicción, ya que los datos son demasiado aleatorios, incluso añadiendo las características de ataque, medio y defensa obtenidas del Dataset de La FIFA, sigue presentándose un sesgo alto en los resultados, esto nos indica que el cambio de entrenador en realidad no está generando un efecto importante en un equipo, de ser el caso real, habría una fuerte correlación entre las variables afines del antes y después, por ejemplo el promedio de victorias en los 5 partidos anteriores al cambio estaría relacionada con el promedio de victorias de los 5 partidos posteriores al cambio, ya que en esta última variable se notaria un resultado similar pero levemente incrementado, pero desafortunadamente no es el caso.

Para realizar predicciones más acertadas en los modelos de aprendizaje supervisado, se requieren variables adicionales al promedio de goles hechos y recibidos anteriores al cambio de entrenador o el promedio de victorias, pues estos datos no tienen encueta factores como la experiencia del entrenador o la plantilla del equipo que realiza el cambio de entrenador, se requeriría variables relacionadas con características de los jugadores, estilos de juego de los entrenadores, táctica y estrategia de juego, incluso rendimiento de los jugadores por individual bajo condiciones de local y visitante.

6. Conclusiones y trabajos futuros

Los objetivos propuestos para este proyecto fueron completados, iniciado por la recolección de datos de diferentes ligas profesionales de primera división, masculinas y femeninas incluyendo datos referentes a los directores técnicos de cada equipo en el tiempo que estuvieron activos. Seguidamente se realizó un proceso de análisis estadístico, que permitió observar el impacto del cambio de entrenador en varias características de los Datasets obtenidos, por ejemplo, el promedio de goles hechos y recibidos, antes y después del cambio de entrenador a corto plazo. Finalmente se aplicaron 4 técnicas de aprendizaje supervisado al conjunto de datos de las ligas europeas masculinas en busca de predecir el promedio de goles y número de puntos generados posterior al cambio de entrenador a corto plazo, a continuación, se describen las conclusiones para los objetivos específicos.

6.1 Objetivo específico número 1

Se recolectaron variedad de datos sobre los encuentros las ligas profesionales masculinas y femeninas de futbol junto con los entrenadores de cada equipo de las ligas seleccionadas, en total se recogieron 38 Datasets bases, con una cantidad de 246334 estancias sumando las ligas europeas masculinas, ligas femeninas y ligas latinoamericanas, y los directores técnicos para todos los equipos de las ligas, incluyendo datos recolectados de los video juegos de LAFIFA. Esta cantidad de datos sirvió como base principal para el análisis estadístico y la construcción de los modelos de Machine Learning.

6.2 Objetivo específico número 2

Gracias a todo lo anterior, se puede interpretar que existe un efecto positivo en el rendimiento de los equipos deportivos de fútbol profesional masculinos y femeninos posterior al cambio de entrenador a corto plazo; En primer lugar, las ligas femeninas poseen una media más alta en las variables estudiadas específicamente en las relacionadas con el promedio de goles hechos y recibidos antes y después del cambio de entrenador, también en la cantidad de puntos hechos antes y después del cambio y finalmente en el porcentaje de victorias antes y después del cambio de entrenador, indicando que en promedio las ligas femeninas generan y reciben más goles también generan más puntos pero reciben más goles.

En cada par de variables afines o de la misma categoría (antes y después), se encontró una diferencia de medias y medianas que aumenta posterior al cambio de entrenador, sin importar si la liga es masculina o femenina lo cual indica que el efecto de cambio de entrenador en promedio tiende a generar un impacto beneficioso en los equipos deportivos de fútbol, ¿Qué tan beneficioso?, para cada variable el cambio es diferente pero siempre en aumento cuando se realiza el cambio de entrenador, por ejemplo en el par de variables relacionadas al promedio de goles hechos antes y después del cambio

a corto plazo (5 encuentros), el porcentaje de incremento para cada variable es de 4.87%, 5.88% y 5.55% (Esta fue la categoría con los valores de incremento más bajos) para las ligas europeas masculinas, europeas femeninas y latinoamericanas respectivamente.

Las ligas femeninas son las más afectada ya que incremento su promedio de goles a más de un 5%. Sin embargo, las ligas europeas y latinoamericanas masculinas tienen prácticamente el mismo incremento. Las únicas variables que recibieron un decremento de cifras posterior al cambio de entrenador fueron las variables relacionadas con los goles recibidos antes y después del cambio de entrenador y el porcentaje de derrotas antes y después del cambio de entrenador. Para las ligas en general, la variable de goles recibidos tuvo un decremento de aproximadamente 13%, siendo las ligas latinoamericanas la categoría con la cifra de incremento más grande (13.4%), mientras que para la variable referente al porcentaje de derrotas el decremento fue de 17% 12% y 18% para las ligas europeas masculinas, ligas femeninas y ligas latinoamericanas respectivamente (El decremento para esta variable es algo positivo pues está indicando que al cambiar de entrenador los equipos reciben menos goles).

La teoría del chivo expiatorio (The ritual scapegoating no-way causality theory), sugiere que el cambio de entrenador es tan solo una fachada, en la cual se cambia al director deportivo con el fin de calmar a los fanáticos y redirigir la culpa del mal desempeño al entrenador, según los análisis esta es la teoría del cambio de entrenador más acertada, ya que los resultados de este trabajo aunque demuestran que tanto para las ligas masculinas como para las femeninas existe un beneficio de gol, victorias y rendimiento en general, esto ocurre solo en promedio pues los datos demuestran una alta dispersión, en otras palabras no se asegura que un equipo deportivo presente los beneficios mencionados por el hecho de cambiar de entrenador, por esta razón este trabajo apoya la teoría del chivo expiatorio extendiendo su explicación también a las ligas femeninas.

6.3 Objetivo específico número 3

Tras el análisis de los modelos de aprendizaje supervisado aplicado para buscar predecir las variables relacionadas con el número de puntos obtenidos y promedio de goles marcados por un equipo posterior al cambio de entrenador, se determinó que los modelos que generan el mejor resultado para predecir el número de puntos fueron 3 modelos, el DecisionTree, el AdaBoost y el GradientBoosting con un MAE de 2.40, y para la predicción de la variable referente al promedio de goles posterior al cambio de entrenador el modelo que genero los mejores resultados fue el modelo GradientBoosting, con un MAE de 0.46, en ambos casos el GradientBoosting fue el modelo con mejores resultados aunque no genero una gran diferencia en comparación con los demás modelos. La variable más relevante en los dos casos fue la variable “puntos_actuales”, indicando que la cantidad de puntos que posee un equipo es de entre todas las variables que más influirá en los modelos. Para complementar lo anterior y según los resultados obtenidos en los modelos de aprendizaje supervisado se puede concluir que para realizar predicciones más acertadas en los modelos, se requieren

variables adicionales al promedio de goles hechos y recibidos anteriores al cambio de entrenador o el promedio de victorias, pues estos datos no tienen encueta factores como la experiencia del entrenador o la plantilla del equipo que realiza el cambio de entrenador, se requeriría variables relacionadas con características de los jugadores, estilos de juego de los entrenadores, táctica y estrategia de juego, incluso rendimiento de los jugadores por individual bajo condiciones de local y visitante, razón por la cual no es posible con las variables y modelos utilizados obtener un resultado capaz de determinar el resultado que obtendrá un club deportivo si decide cambiar su director deportivo en un punto dado.

6.4 Dificultades e imprevistos en la realización del proyecto

1. El proyecto presento algunos desafíos como la obtención de datos de calidad, ya que, en realidad, no existían Datasets que contuvieran la información que se requería en este proyecto, como los resultados de los encuentros y los directores deportivos vigentes en ese encuentro particular, por esta razón se decidió realizar Webscraping en dos páginas para poder obtener los Datos, las páginas fueron “livefutbol.com” y “fifaindex.com”. Estas páginas presentan variedad de datos sobre clubes deportivos y entrenadores al alcance de todas las personas, por otro lado, estos datos no están ordenados en la necesidad que requería este proyecto.

2. La obtención de los datos tomo tiempo debido a que se realizó WebScraping a cada liga por separado, y para cada liga se obtuvo una lista de directores técnicos, por lo que en total se obtuvieron 38 Datasets, Realizar el WebScraping para la obtención de todos los datos conlleva alrededor de 1 semana, si se ejecuta cada programa de WebScraping consecutivamente de uno en uno (Ver Anexos.1.1).

3. Las ligas femeninas en general han tenido menos fama que las ligas masculinas, en los estudios realizados no se encontró registro de teorías que hayan utilizado ligas femeninas de ninguna división, y encontrar datos sobre las ligas femeninas fue más difícil dado que hay pocos registros. Como se observó en la Fase 1, las ligas femeninas poseen registros recientes no tan antiguos como los de las ligas masculinas. De las ligas femeninas seleccionadas, la liga con registros más antiguos fue la Bundesliga con datos desde la temporada (1997/1998), mientras que, para las ligas masculinas, la liga con los registros más antiguos fue la Premier League con datos desde la temporada (1888/1889).

6.5 Aporte a la comunidad y trabajos futuros

Este proyecto se realizó acumulando una gran cantidad de datos de varias ligas masculinas y femeninas profesionales de fútbol, incluyendo ligas latinoamericanas, también se acumularon datos sobre los directores deportivos para todos los equipos de las ligas mencionadas. La cantidad de datos recolectados pueden ser utilizados para otros proyectos que estén relacionados con el fútbol o entrenadores deportivos. En total se obtuvieron 38 Datasets, 19 Datasets con datos relacionados a los partidos jugados de las ligas mencionadas, incluyendo 9 ligas europeas femeninas, y 19 Datasets con la lista de los entrenadores para cada equipo (ver Anexos 1.1.)

El análisis estadístico realizado en las ligas masculinas y femeninas complementa y apoya la teoría del chivo expiatorio (The ritual scapegoating no-way causality theory), sin embargo, el análisis fue realizado a corto plazo (5 encuentros posterior al cambio de entrenador), por lo que se podría extender el estudio a mediano y largo plazo en las ligas femeninas profesionales.

Los modelos de Machine Learning que se obtuvieron pueden ser mejorados incluyendo más variables determinantes relacionadas con las alineaciones utilizadas por los directores técnicos o información sobre los jugadores, ya que se determinó que no se obtienen resultados muy acertados si solo se utilizan características referentes al cambio de entrenador.

7. Referencias

(s.f.). Obtenido de https://www.sngular.com/es/data-science-crisp-dm-metodologia/

90min. (s.f.). *90min*. Obtenido de 90min.com: https://www.90min.com/es/posts/paises-mejor-futbol-femenino-mundo

Amat, J. (2020). Gradient Boosting con Python. Obtenido de https://www.cienciadedatos.net/documentos/py09_gradient_boosting_python.html

ANDES, C.-U. D. (2017). Perfil Alianza Caoba Reporte Tecnico. *Perfil Alianza Caoba Reporte Tecnico*, 41.

As. (s.f.). *As*. Obtenido de As.com: https://colombia.as.com/colombia/2022/01/25/futbol/1643146876_974090.html

Audas, R., Dobson, S., & Goddard, J. (1997). Team performance and managerial change in the English Football League. *Economic Affairs, 17*, 30-36. doi:https://doi.org/10.1111/1468-0270.00039

Audas, R., Dobson, S., & Goddard, J. (2002). The impact of managerial change on team performance in professional sports. *Journal of Economics and Business, 54*, 633-650. doi:https://doi.org/10.1016/S0148-6195(02)00120-0

Besters, L., van Ours, J., & van Tuijl , M. (2016). Effectiveness of In-Season Manager Changes in English Premier League Football. *Football. De Economist, 164*, 335-356. doi:https://doi.org/10.1007/s10645-016-9277-0

d'Addona, S., & Kind, A. (2012). Forced Manager Turnovers in English Soccer Leagues: A Long-Term Perspective. *Journal of Sports Economics, 15*, 150-189. doi:https://doi.org/10.1177/1527002512447803

Eitzen, D., & Yetman, N. (1972). Managerial Change, Longevity, and Organizational Effectiveness. *Administrative Science Quarterly, 17*, 110-116. doi:https://doi.org/10.2307/2392099

fifaindex. (s.f.). *fifaindex*. Obtenido de fifaindex.com: https://www.fifaindex.com/

Flint, S., Plumley, D., & Wilson, R. (2016). You're getting sacked in the morning: managerial change in the English Premier League. *Marketing Intelligence &, 34*, 223-235. doi:https://doi.org/10.1108/MIP-09-2014-0189

Gamson, W., & Scotch, N. (Julio de 1964). Scapegoating in Baseball. *American Journal of Sociology, 70*, 69-72. doi:https://doi.org/10.1086/223739

Grusky. (1963). Managerial succession and organizational effectiveness. *American Journal, 26*, 21-31. doi:https://doi.org/10.1086/223507

Hentschel, S., Muehlheusser, G., & Sliwka, D. (2012). The impact of managerial change on performance: The role of team heterogeneity. *Journal Sports Economics*. doi:https://doi.org/10.2139/ssrn.2158294

Hernández G, C., & Dueñas R., M. (2009). Hacia una metodología de gestión del conocimiento basada en minería de datos. *Hacia una metodología de gestión del conocimiento basada en minería de datos*, 95.

Heuer, A., Müller, C., Rubner, O., Hagemann, N., & Strauss, B. (2011). Usefulness of Dismissing and Changing the Coach in Professional Soccer. *PloS one*. doi:https://doi.org/10.1371/journal.pone.0017664

learn, s. (s.f.). *scikit learn.* Obtenido de scikit-learn.org: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.r2_score.html#sklearn.metrics.r2_score

livefutbol. (s.f.). *livefutbol*. Obtenido de livefutbol.com: https://www.livefutbol.com/

Pedregosa, F. a. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research, 12*, 2825-2830.

Peñas-Lago. (2011). Coach Mid-Season Replacement and Team Performance in Professional Soccer. *Journal of human kinetics, 28*, 115-122. doi:https://doi.org/10.2478/v10078-011-0028-7

Soebbing, B., Wicker, P., & Weimar, D. (2015). The Impact of Leadership Changes on Expectations of Organizational Performance. *Journal of Sport Management, 29*, 485-497. doi:https://doi.org/10.1123/jsm.2014-0089

Wagner, S. (2010). Managerial succession and organizational performance—Evidence from the German Soccer League. *Managerial and Decision Economics, 31*, 415-430. doi:https://doi.org/10.1002/mde.1495

Anexos

1.1. Repositorio en GitHub

Para obtener los programas utilizados en este trabajo y los Datasets obtenidos con dichos programas por favor utilizar el siguiente link:

https://github.com/AndreySuavita/AndreySuavita-Analisis-del-impacto-de-cambio-de-entrenador-en-ligas-de-futbol

En el repositorio se encontrarán las siguientes carpetas:

- **Programas_python:**
  Contiene los archivos.py utilizados para extraer los datos de fútbol de las páginas correspondientes y los programas que fusionan dichos Datasets.
- **Dataset_fifa:**
  Contiene un Dataset con información obtenida de la página “fifaindex.com” con información sobre características de los equipos deportivos.
- **datasets_results:**
  Contiene todos los Datasets obtenidos por los programas de Python de las ligas masculinas de fútbol.
- **datasets_results_f:**
  Contiene todos los Datasets obtenidos por los programas de Python de las ligas femeninas de fútbol.
- **datasets_unificados:**
  Contiene 3 Datasets, el primero es el resultado de la fusión entre todas ligas masculinas europeas, el segundo es el resultado de la fusión de todas las ligas femeninas, y el tercero es el resultado de la fusión de todas las ligas latinoamericanas masculinas.
- **datasets_unificados_limpios:**
  Contiene los Datasets unificados limpios y transformados para realizar el análisis, estadístico y el Machine Learning y también posee la adición de las características de los equipos obtenidas por el Dataset de LAFIFA en los equipos de las ligas europeas.
- **managers:**
  Contiene los Datasets con información de los entrenadores de cada una de las ligas masculinas, y la unificación de dichos Datasets.
- **managers_f:**
  Contiene los Datasets con información de los entrenadores de cada una de las ligas femeninas, y la unificación de dichos Datasets.

7.1 Archivos en Google Colab

Para visualizar y descargar los NooteBook utilizados en Google Colab utilizados en este trabajo para realizar en análisis estadístico y el machine learning por favor utilizar el siguiente link.

https://drive.google.com/drive/folders/1C2aZxgwl7JIUwwdLpHuR01FRcr8JDRuD?usp=sharing

La carpeta de Google Colab contiene 7 archivos enumerados con los procesos realizados para la trasformación, análisis estadístico y construcción de modelos de Machine Learning:

1. **Limpieza_y_transformacion_de_datasets_unificados:**
   En este cuaderno se realizó el proceso de limpieza y transformación para los Datasets de las ligas europeas masculinas, ligas latinoamericanas masculinas y ligas femeninas.
2. **Estadistica_descriptiva_datasets_limpios:**
   En este cuaderno se realizó el proceso de estadística descriptiva para la determinación del efecto de cambio de entrenador en cada uno de los Datasets trasnformados.
3. **Web_Scraping_fifa:**
   En este cuaderno se obtuvieron datos del ataque, medio y defensa de los equipos deportivos masculinos de la página “fifaindex.com” desde el año 2005 hasta el año 2022.
4. **Fusion_fifa_entrenadores:**
   En este cuaderno se realizó la fusión del Dataset obtenido en el cuaderno anterior de LAFIFA y el Dataset de entrenadores de las ligas europeas masculinas.
5. **Combinar_fifa_dataset_limpio:**
   En este cuaderno se realizó la fusión entre las características de LAFIFA y los Datasets de las ligas europeas masculinas.
6. **Seleccion_caracteristicas_datasets_unificados:**
   En este cuaderno se aplicaron técnicas de aprendizaje no supervisado utilizando los Datasets de las ligas europeas masculinas, con el fin de seleccionar características o variables de utilidad para ser utilizadas en el proceso de predicción.
7. **Modelos_prediccion_goles_y_puntos:**
   En este cuaderno se realizaron los modelos de aprendizaje supervisado para la predicción del promedio de goles hechos y puntos hechos después de que se efectué el cambio de entrenador en algún equipo deportivo de fútbol.

7.2 Resumen de resultados parte 1

- ***Parte 1: Resultado de modelos de regresión lineal***

A continuación, se pueden observar las gráficas de las variables utilizadas en la regresión lineal junto con los resultados de cada modelo de regresión lineal.

- **Resultados ligas europeas masculinas**

Modelo 1	Europa Ligas masculinas
Variable independiente	m_goles_hechos_antes
Variable dependiente	m_goles_hechos_despues
Ejemplos para entrenamiento	304
Ejemplos para test	76
Coeficiente 1	0.578301
Intersección	0.51694431
Error cuadrático medio	0.06
Error absoluto medio	0.19
R2	0.16
Modelo 2	Europa Ligas masculinas
Variable independiente	m_victorias_antes
Variable dependiente	m_victorias_despues
Ejemplos para entrenamiento	291
Ejemplos para test	73
Coeficiente 1	0.366893
Intersección	0.18628805
Error cuadrático medio	0.01
Error absoluto medio	0.07
R2	0.02
Modelo 3	Europa Ligas masculinas
Variable independiente	m_goles_recibidos_antes
Variable dependiente	m_goles_recibidos_despues
Ejemplos para entrenamiento	300
Ejemplos para test	75
Coeficiente 1	0.438593
Intersección	0.80121463
Error cuadrático medio	0.07
Error absoluto medio	0.21
R2	0.24
Modelo 4	Europa Ligas masculinas

Variable independiente	m_puntos_hechos_antes
Variable dependiente	m_puntos_hechos_despues
Ejemplos para entrenamiento	298
Ejemplos para test	75
Coeficiente 1	0.301747
Intersección	3.66170996
Error cuadrático medio	2.19
Error absoluto medio	1.13
R2	0.08

Tabla 41. Resumen de resultados de modelos de regresión lineal Ligas europeas masculinas. Elaboración propia.

- **Ligas Latinoamericanas masculinas**

Modelo 1	Latinoamérica Ligas masculinas
Variable independiente	m_goles_hechos_antes
Variable dependiente	m_goles_hechos_despues
Ejemplos para entrenamiento	93
Ejemplos para test	24
Coeficiente 1	0.63101
Intersección	0.43823681
Error cuadrático medio	0.05
Error absoluto medio	0.16
R2	0.01
Modelo 2	Latinoamérica Ligas masculinas
Variable independiente	m_victorias_antes
Variable dependiente	m_victorias_despues
Ejemplos para entrenamiento	92
Ejemplos para test	24
Coeficiente 1	0.545025
Intersección	0.16337775
Error cuadrático medio	0.006
Error absoluto medio	0.06
R2	0.31
Modelo 3	Latinoamérica Ligas masculinas
Variable independiente	m_goles_recibidos_antes
Variable dependiente	m_goles_recibidos_despues

Ejemplos para entrenamiento	94
Ejemplos para test	24
Coeficiente 1	0.589992
Intersección	0.43521369
Error cuadrático medio	0.05
Error absoluto medio	0.16
R2	0.31
Modelo 4	Latinoamérica Ligas masculinas
Variable independiente	m_puntos_hechos_antes
Variable dependiente	m_puntos_hechos_despues
Ejemplos para entrenamiento	89
Ejemplos para test	23
Coeficiente 1	0.542455
Intersección	2.52860949
Error cuadrático medio	0.97
Error absoluto medio	0.78
R2	0.43

Tabla 42. Resumen de resultados de modelos de regresión lineal Ligas europeas masculinas. Elaboración propia.

- **Ligas femeninas**

Modelo 1	Europa Ligas femeninas
Variable independiente	m_goles_hechos_antes
Variable dependiente	m_goles_hechos_despues
Ejemplos para entrenamiento	108
Ejemplos para test	28
Coeficiente 1	0.644828
Intersección	0.55807469
Error cuadrático medio	0.47
Error absoluto medio	0.52
R2	0.008
Modelo 2	Europa Ligas femeninas
Variable independiente	m_victorias_antes
Variable dependiente	m_victorias_despues
Ejemplos para entrenamiento	92
Ejemplos para test	24
Coeficiente 1	0.545025
Intersección	0.16337775

Error cuadrático medio	0.006
Error absoluto medio	0.06
R2	0.31
Modelo 3	Europa Ligas femeninas
Variable independiente	m_goles_recibidos_antes
Variable dependiente	m_goles_recibidos_despues
Ejemplos para entrenamiento	113
Ejemplos para test	29
Coeficiente 1	0.471353
Intersección	0.83473654
Error cuadrático medio	0.33
Error absoluto medio	0.46
R2	-0.30
Modelo 4	Europa Ligas femeninas
Variable independiente	m_puntos_hechos_antes
Variable dependiente	m_puntos_hechos_despues
Ejemplos para entrenamiento	116
Ejemplos para test	29
Coeficiente 1	0.539062
Intersección	3.45075506
Error cuadrático medio	9.35
Error absoluto medio	2.51
R2	0.33

Tabla 43. Resumen de resultados de modelos de regresión lineal Ligas europeas masculinas. Elaboración propia.

7.3 Resumen de resultados parte 2

Se aplicaron cuatro modelos de aprendizaje supervisado (RandomForests, DecisionTree, AdaBoost, GradientBoosting) al conjunto de datos con el objetivo de tratar de predecir las variables llamadas "puntos_hechos_despues_de_cambio_entrenador_1_5" y "promedio_goles_hechos_despues_de_cambio_entrenador_1_5".

Las características utilizadas en los modelos para predecir la variable "puntos_hechos_despues_de_cambio_entrenador_1_5" son las siguientes:

Variable	Descripción
jornada	Número de jornada de la temporada en la que se realizó el cambio de entrenador.
puntos_actuales	Puntos que poseía el equipo en el momento del cambio de entrenador.
puntos_hechos_antes_de_cambio_entrendor_1_5	Puntos generados en los últimos 5 encuentros antes del cambio de entrenador.
ataque	Promedio de ataque del equipo que realiza el cambio de entrenador.

Tabla 44. Características utilizadas en los modelos (variable: puntos_hechos_despues_de_cambio_entrenador_1_5). Elaboración propia.

Resumen de resultados e hiperparámetros utilizados en los modelos:

Modelo 1	RandomForests
Variable Objetivo	puntos_hechos_despues_de_cambio_entrendor_1_5
n_estimators	1024
Criterion	mae
max_depth	Default (None)
random_state	0
MAE Obtenido	2.54
MSE Obtenido	9.59
variable más relevante	ataque (0,29)
variable menos relevante	jornada (0,21)

Tabla 45. Resultados del modelo 1(variable: puntos_hechos_despues_de_cambio_entrendor_1_5). Elaboración propia.

Modelo 2	DecisionTree
Variable Objetivo	puntos_hechos_despues_de_cambio_entrendor_1_5
max_depth	7
Criterion	mae
MAE Obtenido	2.40
MSE Obtenido	8.86
variable más relevante	puntos_actuales (0,52)
variable menos relevante	puntos_hechos_antes_de_cambio_entrenador_1_5 (0,05)

Tabla 46. Resultados del modelo 2 (variable: puntos_hechos_despues_de_cambio_entrendor_1_5). Elaboración propia.

Modelo 3	AdaBoost
Variable Objetivo	puntos_hechos_despues_de_cambio_entrendor_1_5
n_estimators	16
Criterion	mae
learning_rate	Default (0,01)
loss	Default (linear)
random_state	0
MAE Obtenido	2.4
MSE Obtenido	8.67
variable más relevante	puntos_actuales (0,35)
variable menos relevante	ataque (0,13)

Tabla 47. Resultados del modelo 3 (variable: puntos_hechos_despues_de_cambio_entrendor_1_5). Elaboración propia.

Modelo 4	GradientBoosting
Variable Objetivo	puntos_hechos_despues_de_cambio_entrendor_1_5
n_estimators	1024
Criterion	mae
learning_rate	0.01
loss	absolute_error
random_state	0
MAE Obtenido	2.4
MSE Obtenido	8.76
variable más relevante	puntos_actuales(0,29)
variable menos relevante	puntos_hechos_antes_de_cambio_entrenador_1_5 (0.15)

Tabla 48. Resultados del modelo 4 (variable: puntos_hechos_despues_de_cambio_entrendor_1_5). Elaboración propia.

Las características utilizadas en los modelos para predecir la variable "promedio_goles_hechos_despues_de_cambio_entrenador_1_5" son las siguientes:

Variable	Descripción
jornada	Número de jornada de la temporada en la que se realizó el cambio de entrenador.
puntos_actuales	Puntos que poseía el equipo en el momento del cambio de entrenador.
promedio_goles_hechos_antes_de_cambio_entrenador_1_5	Promedio de goles hechos en los 5 encuentros anteriores al cambio de entrenador.
ataque	Promedio de ataque del equipo que realiza el cambio de entrenador.

Tabla 49. Características utilizadas en los modelos (variable: promedio_goles_hechos_despues_de_cambio_entrenador_1_5). Elaboración propia.

Resumen de resultados e de hiperparámetros utilizados en los modelos:

Modelo 1	RandomForests
Variable Objetivo	promedio_goles_hechos_despues_de_cambio_entrendor_1_5
n_estimators	1024
Criterion	mae
max_depth	Default (None)
random_state	0
MAE Obtenido	0,50
MSE Obtenido	0.39
variable más relevante	ataque (0.31)
variable menos relevante	jornada (0,20)

Tabla 50. Resultados del modelo 1 (variable: promedio_goles_hechos_despues_de_cambio_entrendor_1_5). Elaboración propia.

Modelo 2	DecisionTree
Variable Objetivo	promedio_goles_hechos_despues_de_cambio_entrendor_1_5
max_depth	5
Criterion	mae
MAE Obtenido	0.47
MSE Obtenido	0.37
variable más relevante	promedio_goles_hechos_antes_de_cambio_entrenador_1_5 (0,41)
variable menos relevante	jornada (0,07)

Tabla 51. Resultados del modelo 2 (variable: promedio_goles_hechos_despues_de_cambio_entrendor_1_5). Elaboración propia.

Modelo 3	AdaBoost
Variable Objetivo	promedio_goles_hechos_despues_de_cambio_entrendor_1_5
n_estimators	4
Criterion	mae
learning_rate	Default (0,01)
loss	Default (linear)
random_state	0
MAE Obtenido	0.48
MSE Obtenido	0.36
variable más relevante	puntos_actuales (0,33)
variable menos relevante	jornada (0,20)

Tabla 52. Resultados del modelo 3 (variable: promedio_goles_hechos_despues_de_cambio_entrendor_1_5). Elaboración propia.

Modelo 4	GradientBoosting
Variable Objetivo	promedio_goles_hechos_despues_de_cambio_entrendor_1_5
n_estimators	1024
Criterion	mae
learning_rate	0,01
loss	absolute_error
random_state	0
MAE Obtenido	0,46
MSE Obtenido	0.34
variable más relevante	puntos_actuales (0,29)
variable menos relevante	jornada (0,16)

Tabla 53. Resultados del modelo 4 (variable: promedio_goles_hechos_despues_de_cambio_entrendor_1_5). Elaboración propia.