Universidad de Valladolid

Máster universitario

Ingeniería Informática

Trabajo Fin de Máster

Comparación y evaluación de diferentes técnicas de IA para un modelo de rotación de empleados en empresas aplicado a equipos de fútbol

Realizado por **José María Lozano Olmedo**

✠✠✠

Universidad de Valladolid

19 de junio de 2024

Tutor: Joaquín Adiego Rodríguez y Diego Rafael Llanos Ferraris

Universidad de Valladolid

Máster universitario en Ingeniería Informática

D. Joaquín Adiego Rodríguez y Diego Rafael Llanos Ferraris, profesor del departamento de DEPARTAMENTO DEL TUTOR, área de AREA_CONOCIMIENTO DEL TUTOR.

Expone:

Que el alumno D. José María Lozano Olmedo, ha realizado el Trabajo final de Máster en Ingeniería Informática titulado "Comparación y evaluación de diferentes técnicas de IA para un modelo de rotación de empleados en empresas aplicado a equipos de fútbol".

Y que dicho trabajo ha sido realizado por el alumno bajo la dirección del que suscribe, en virtud de lo cual se autoriza su presentación y defensa.

En Valladolid, 19 de junio de 2024

Vº. Bº. del Tutor:

D. nombre tutor

Vº. Bº. del co-tutor:

D. nombre co-tutor

Resumen

En los últimos años, la inteligencia artificial ha experimentado un exponencial crecimiento, revolucionando diferentes sectores con sus innovadoras ideas. Este crecimiento tecnológico no ha pasado desapercibido en el ámbito deportivo, donde la inteligencia artificial se está incorporando cada vez más para optimizar el rendimiento de los equipos y atletas.

Este trabajo se centra en evaluar diversas técnicas de inteligencia artificial para predecir el rendimiento de equipos de fútbol mediante el análisis de la rotación de jugadores. El objetivo del proyecto es detectar cómo las diferentes estrategias de rotación aplicadas por los equipos afectan al desempeño del equipo y cómo la inteligencia artificial puede realizar predicciones en base a ellas para ayudar a optimizar estas estrategias y detectar cuáles son las mejores. Este trabajo abarca desde la recopilación y el análisis de datos asociados a las ligas seleccionadas, la creación de modelos de inteligencia artificial y la evaluación de su eficacia. Se pretende que con este proyecto se puedan obtener resultados significativos para optimizar la gestión de equipos de fútbol y así poder facilitar el trabajo a sus dirigentes.

Descriptores

Fútbol, jugadores, inteligencia artificial, predicciones, rotación, rendimiento

Abstract

In recent years, artificial intelligence has experienced exponential growth, revolutionizing various sectors with its innovative ideas. This technological advancement has not gone unnoticed in the sports field, where artificial intelligence is increasingly being integrated to optimize team and athlete performance.

This study focuses on evaluating various artificial intelligence techniques to predict the performance of football teams through the analysis of player rotation. The objective of the project is to detect how different rotation strategies applied by teams impact their performance, and how artificial intelligence can make predictions based on these strategies to optimize them and identify the best ones. This work ranges data collection and analysis from selected leagues, the development of artificial intelligence models, and the evaluation of their effectiveness. The goal is to achieve significant results with this project to optimize the management of football teams and thereby facilitate the work of their leaders.

Keywords

Football, players, artificial intelligence, predictions, rotation, performance

Índice general

Índice de figuras

Índice de tablas

1: Introducción

1.1. Contexto

Este proyecto se desarrolla como el Trabajo de Fin de Máster del Máster en Ingeniería Informática No Presencial de la Escuela de Ingeniería Informática de la Universidad de Valladolid y como continuación del trabajo realizado durante la estancia en un GIR para la asignatura de I+D+i.

En los últimos años, el campo de la inteligencia artificial ha sufrido un crecimiento considerable, alterando diferentes aspectos de la sociedad moderna [20]. En este contexto, el deporte, y en especial el fútbol, no se ha mantenido al margen. La capacidad de la inteligencia artificial para analizar grandes volúmenes de datos y extraer patrones útiles ha encontrado una aplicación cada vez más importante en el ámbito deportivo, proporcionando nuevas herramientas para optimizar el rendimiento de los equipos y la toma de decisiones por parte de los directivos y entrenadores [23].

El fútbol, es algo más que un simple juego, se ha convertido en un fenómeno global que supera todas las fronteras. Los clubes de fútbol son empresas con mucho dinero y todo lo relacionado con este deporte, de manera general, mueve grandes cantidades de dinero. En este contexto altamente competitivo, la presión por obtener resultados positivos es máxima, tanto en términos deportivos como financieros [18].

Un claro ejemplo de esto es el fichaje de Neymar por el Paris Saint-Germain (PSG) en 2017, que se firmó por la desorbitada cantidad de 222 millones de euros. Este traspaso no solo rompió récords, sino que también destacó cómo los grandes clubes están dispuestos a invertir grandes cantidades para atraer a los mejores jugadores y a su vez, generar ingresos mediante la venta de camisetas, patrocinadores y entradas a los partidos. La presencia de estrellas de este nivel en los equipos aumenta el valor comercial de los clubes, atrayendo aún más patrocinadores, haciendo mejores los acuerdos de *marketing* y por lo tanto mejorando el estado financiero de los equipos [11].

Por otro lado, los derechos de televisión son otra fuente crucial de ingresos en el fútbol. Las ligas más importantes, como la Premier League inglesa, reciben miles de millones de libras por la transmisión de sus partidos. Estos contratos televisivos proporcionan la

capacidad a los clubes de repartir grandes sumas de dinero entre ellos, lo que permite financiar fichajes y salarios enormes. El torneo británico concedió sus derechos televisivos en la temporada 2022/2023 por todo el mundo por un valor total de 8.136 millones de libras [38].

Además, los estadios modernos se han convertido en auténticas obras de ingeniería. El Santiago Bernabéu del Real Madrid, por ejemplo, está siendo remodelado con una inversión cercana a los 1950 millones de euros. Esta obra no solo incrementará considerablemente la experiencia de los aficionados, sino que también permitirá al club generar ingresos adicionales a través de eventos no deportivos como conciertos y nuevas áreas comerciales, como ya está sucediendo. Así, el fútbol se está convirtiendo en una industria multimillonaria que atrae inversiones de todo el mundo, consolidándose como el principal espectáculo mundial [36].

En este escenario, la gestión eficiente de los recursos humanos, como en este caso los jugadores, se ha vuelto prioritaria para el éxito de un equipo. Los entrenadores y directivos se enfrentan al desafío de optimizar el rendimiento de sus jugadores tratando de minimizar el riesgo de lesiones y el cansancio físico. La inteligencia artificial ofrece herramientas potentes para abordar este desafío, permitiendo el análisis de datos relacionados con el estado físico de los jugadores, el rendimiento en partidos anteriores, las lesiones previas y otros factores relevantes.

La gestión de la rotación de jugadores es uno de los aspectos más críticos de la estrategia de un equipo a lo largo de una temporada y que tiene una mayor repercusión sobre su éxito. La inteligencia artificial puede ayudar a los entrenadores a tomar decisiones documentadas sobre cuándo dar descanso a un jugador, cuándo alinear a un futbolista, qué cambios realizar y cómo mantener un equilibrio entre la competitividad y la salud de la plantilla.

Los equipos de fútbol utilizan el análisis de datos para gestionar eficientemente a sus jugadores y optimizar su rendimiento en el campo, así como para reducir el riesgo de lesiones. Mediante tecnologías avanzadas y técnicas de última generación, los clubes recopilan un amplio campo de datos, incluyendo métricas de rendimiento físico como la distancia recorrida, velocidad, y cargas de trabajo. Estos datos permiten a los entrenadores y al personal médico controlar la condición física que tienen los jugadores en tiempo real, identificar patrones de fatiga, y ajustar los entrenamientos y tiempos de juego según como corresponda en base a los datos recopilados.

Un club que aplica esto es el Liverpool FC que emplea el análisis de datos para tomar decisiones justificadas sobre cuándo rotar a sus jugadores, garantizando no sobrepasar sus límites físicos y que se mantengan al máximo rendimiento durante toda la temporada. Este enfoque basado en datos no solo aumenta el rendimiento individual y colectivo del equipo, sino que también ayuda a evitar lesiones, prolongando así la carrera de los jugadores y garantizando su disponibilidad y rendimiento para los partidos clave de la temporada donde es necesario que estén al 100 % [5].

Por lo tanto, el uso de la inteligencia artificial en el fútbol no solo es una oportunidad para mejorar el rendimiento deportivo, sino también una necesidad para mejorar sobre

los rivales en un entorno cada vez más competitivo y exigente. Los equipos que puedan lograr aprovechar e incorporar de manera efectiva estas herramientas tendrán una ventaja considerable sobre el resto para conseguir sus objetivos deportivos y financieros.

1.2. Motivación

El crecimiento en los últimos años de la inteligencia artificial ha despertado un interés en su aplicación en diversos campos como en el deporte. En el ámbito del fútbol, la capacidad de utilizar la inteligencia artificial para analizar datos complejos y tomar decisiones estratégicas concretas en base a ellos, ofrece un gran potencial para incrementar el rendimiento de los equipos. Esta motivación se debe a la necesidad de los clubes por mantenerse competitivos en un entorno que está en constante evolución, donde la línea entre el éxito y el fracaso cada vez es más estrecha.

La inteligencia artificial ha transformado el ámbito deportivo al proporcionar herramientas avanzadas para el análisis de datos, la mejora del rendimiento y la toma de decisiones justificadas. En el fútbol, la inteligencia artificial se utiliza para analizar enormes volúmenes de datos de partidos, dando la capacidad a los entrenadores de desarrollar tácticas más precisas, adecuadas y adaptadas. Además, la inteligencia artificial optimiza la gestión de la salud de los deportistas al predecir el riesgo de lesiones y proponer programas de entrenamiento personalizados que se adapten a ellos. Fuera del campo, la inteligencia artificial cada vez mejora más la experiencia de los aficionados. La inteligencia artificial está revolucionando la forma en que los equipos y atletas se preparan y compiten, elevando el deporte de alto nivel a nuevos escalones [2].

El proyecto aparece como respuesta al incremento en la demanda de herramientas que permitan a los clubes mejorar en la gestión de sus recursos humanos, en concreto de sus jugadores. La inteligencia artificial tiene la capacidad de analizar grandes cantidades de datos sobre los jugadores y equipos, detectando patrones y tendencias que pueden pasar desapercibidos para las personas. Al integrar estas conclusiones en la toma de decisiones, los equipos pueden mejorar la eficiencia de su rotación de jugadores, incrementando así sus posibilidades de éxito en el campo.

En los últimos años, la demanda de especialistas en análisis de datos en el ámbito deportivo ha aumentado considerablemente debido a la creciente importancia de los datos para optimizar el rendimiento y la estrategia de los equipos y atletas. Equipos y organizaciones deportivas ahora buscan analistas de datos para analizar métricas de rendimiento, evaluar riesgos de lesiones y desarrollar tácticas más eficientes contra sus rivales. Este crecimiento se debe a los avances tecnológicos y a la comprensión de que el análisis de datos puede generar una ventaja competitiva considerable, incrementando tanto el desempeño individual de los atletas como los resultados colectivos de los equipos [16].

Este proyecto tiene el potencial para marcar un cambio significativo en la manera en que se realiza la gestión deportiva en el fútbol moderno. Al ofrecer a los clubes herramientas

avanzadas de análisis y toma de decisiones, se espera que este proyecto ayude no solo a mejorar los resultados deportivos, sino también a fortalecer la posición competitiva y el rendimiento financiero de los equipos en un mercado cada vez más exigente y competitivo.

1.3. Aplicaciones similares

A continuación, se detallan aplicaciones y proyectos similares a lo que se pretende desarrollar y que pueden servir de referencia.

- **LaLiga Beyond Stats:** esta es una iniciativa de LaLiga que tiene como objetivo emplear las últimas tecnologías, relacionadas con el análisis de datos y la inteligencia artificial, para proporcionar una comprensión más profunda y completa de los partidos. Esta plataforma busca ofrecer a los aficionados, entrenadores, jugadores y clubes herramientas innovadoras para analizar y entender el rendimiento en el fútbol, más allá de las estadísticas habituales, a través de datos en tiempo real y visualizaciones interactivas, proporcionando así un enfoque más inteligente y apasionado hacia el deporte [21].
- **Aplicación de la inteligencia artificial en la Premier League:** esta liga utiliza la inteligencia artificial para determinar las probabilidades de que un equipo gane un partido mediante el análisis de un amplio rango de datos. Estos factores abarcan datos históricos de partidos anteriores, como el rendimiento del equipo en casa y fuera de casa, su posición en la tabla de clasificación, su forma actual y lesiones de jugadores clave entre otros. Además, se tienen en cuenta variables más específicas, como la posesión de balón, los tiros a puerta, las oportunidades creadas y el rendimiento en defensa y en ataque. Estos datos son proporcionados a algoritmos de aprendizaje automático que son capaces de analizar patrones complejos y entrenar modelos predictivos que estiman las probabilidades de que gane el local, empaten o gane el visitante. De esta manera, la inteligencia artificial proporciona una herramienta útil para predecir resultados de partidos de fútbol con un alto grado de precisión, lo que puede ser utilizado por equipos, aficionados y casas de apuestas para tomar decisiones en base a diferentes argumentos [28] [35].
- **Opta:** es una empresa líder en análisis y datos deportivos que es capaz de proporcionar información detallada y estadísticas sobre una amplia gama de eventos deportivos, incluyendo fútbol, *rugby*, *cricket* y otros. Para ello, utiliza tecnologías avanzadas de recopilación y análisis de datos donde recopila datos en tiempo real durante los eventos deportivos y los convierte en información valiosa y estadísticas significativas que son utilizadas por equipos, entrenadores, medios de comunicación y aficionados para analizar de mejor manera el juego y evaluar el rendimiento de los jugadores y equipos. Opta se ha convertido en un recurso importantísimo en el mundo del deporte para análisis de datos y seguimiento de estadísticas [27].
- **Mejora del rendimiento de los jugadores mediante la ciencia de datos en el Liverpool FC:** como se ha comentado previamente, uno de los principales equipos

y pioneros en este área que utiliza la ciencia de datos e inteligencia artificial para mejorar el rendimiento de sus jugadores es el Liverpool FC. Mediante un enfoque avanzado, emplean grandes cantidades de datos, incluyendo métricas de rendimiento físico, táctico y de salud, recogidos durante sus entrenamientos y partidos. Estos datos son analizados para mejorar las estrategias de juego, establecer entrenamientos personalizados y gestionar la carga de trabajo de los jugadores, lo que proporciona una gran ayuda a prevenir lesiones. Además, el análisis de datos permite al cuerpo técnico tomar decisiones razonadas sobre alineaciones y tácticas, ajustando su enfoque basándose en el rendimiento y las características del oponente, lo que contribuye considerablemente al éxito del equipo en la temporada.

1.4. Estructura de la memoria

Este documento se estructura de la siguiente forma:

Capítulo 2 Objetivos del proyecto: en este capítulo se describen los objetivos que se quieren conseguir con la ejecución de este proyecto. Estos se dividen en dos categorías distintas, los objetivos de desarrollo y los objetivos académicos.

Capítulo 3 Conceptos teóricos: en este capítulo se expone una explicación teórica de los conceptos más importantes que se han utilizado para el desarrollo de este proyecto.

Capítulo 4 Técnicas y herramientas: en este capítulo se describen las técnicas utilizadas para la obtención de los datos y las tecnologías utilizadas para el desarrollo del proyecto.

Capítulo 5 Aspectos relevantes del desarrollo del proyecto: en este capítulo se recogen los aspectos más interesantes del desarrollo del proyecto como los detalles sobre las fases de análisis, diseño e implementación y el tratamiento de los datos.

Capítulo 6 Resultados: en este capítulo se exponen los resultados obtenidos y todas las conclusiones que se han extraído en el desarrollo del proyecto y que pueden ayudar a los entrenadores a cómo gestionar los jugadores.

Capítulo 7 Conclusiones y líneas de trabajo futuras: en este capítulo se explican las conclusiones finales adquiridas del proyecto junto a las posibles líneas de trabajo futuras a seguir.

Apéndice Manual de instalación: en este apéndice se describe el repositorio donde se encuentra el código y cómo se puede instalar y desplegar.

2: Objetivos del proyecto

2.1. Introducción

En este capítulo se describen los objetivos que se pretenden conseguir con este proyecto, diferenciando entre objetivos de desarrollo y académicos.

2.2. Objetivos de desarrollo

El principal objetivo de desarrollo es entrenar varios modelos con inteligencia artificial que ayuden a los entrenadores a tomar mejores decisiones sobre qué jugadores utilizar en un partido mediante los datos obtenidos en los partidos anteriores. Para ello, los principales objetivos de desarrollo para este proyecto son:

1. **Obtener los datos de los partidos de fútbol de varias ligas.** Para ello, mediante el *scraping* se extraerán los datos de todos los partidos jugados en diferentes ligas y la información asociada a los jugadores.
2. **Limpiar, transformar y analizar los datos obtenidos.** Se deberán limpiar y transformar los datos obtenidos para que puedan ser utilizados por los modelos que se pretenden crear. Además, se debe realizar un análisis previo sobre los datos para detectar posibles patrones.
3. **Aplicar diferentes modelos de inteligencia artificial con los datos obtenidos.** En este punto, se debe evaluar el rendimiento que tienen los diferentes modelos sobre los datos obtenidos.
4. **Optimizar el rendimiento de los modelos.** Después de entrenar los diferentes modelos sobre los datos, se debe realizar una optimización de sus parámetros para mejorar la precisión obtenida.
5. **Seleccionar los mejores modelos y analizar su rendimiento proporcionado.** Sobre todos los modelos evaluados, se deberán seleccionar los que mejor se comporten y se deberá de analizar que calidad tienen.

6. **Documentar los pasos seguidos en el proyecto.** Se deben documentar todos los pasos seguidos en el proyecto y justificar todas las decisiones tomadas incluyendo los objetivos, métodos y resultados de la investigación y desarrollo. Por otro lado se debe detallar la estructura, funcionamiento y uso del código.

7. **Obtener conclusiones determinantes sobre qué estrategias son mejores.** Al finalizar el proyecto se deben obtener conclusiones que ayuden a los entrenadores a gestionar de mejor manera los jugadores y tras el desarrollo de este proyecto se debe justificar adecuadamente qué estrategias son las mejores.

2.3. Objetivos académicos

Estos objetivos se centran en seguir profundizando en los conocimientos aprendidos en diversas asignaturas de este Máster relacionadas con el *Deep Learning* y el *Big Data* y ponerlos en práctica en un proyecto completo. A continuación se detallan cada uno de estos objetivos:

1. **Aplicar los conocimientos asociados a los pasos de extracción, transformación y carga de los datos.** Se pretende poner en práctica todos los conocimientos asociados al proceso de ETL *(Extract, Transform and Load)* seguido en los proyectos de *Big Data* para estructurar los datos y que puedan ser utilizados por los modelos.

2. **Aplicar las técnicas aprendidas sobre modelos de inteligencia artificial y redes neuronales.** Se pretende seguir profundizando y aplicar los conocimientos adquiridos sobre redes neuronales e inteligencia artificial para que los modelos creados tengan la mayor precisión posible.

3. **Aplicar las mejores prácticas aprendidas para crear gráficos para el análisis de los datos.** Se pretende realizar un análisis de los datos obtenidos para detectar patrones que puedan ayudar a los entrenadores en la toma de decisiones. En este análisis, se debe perseguir que los gráficos creados cumplan con las mejores prácticas que se han aprendido para la creación de interfaces gráficas para el análisis de datos.

3: Conceptos teóricos

3.1. Introducción

Las técnicas de inteligencia artificial abarcan una amplia gama de metodologías y enfoques. A continuación se detallan las técnicas más importantes que se han utilizado en este proyecto y sus conceptos teóricos.

3.2. Introducción a la inteligencia artificial

La inteligencia artificial (IA) es un campo de la informática que se dedica a la creación de sistemas que son capaces de realizar tareas que generalmente requieren capacidades humanas. Estas tareas abarcan multitud de acciones desde el reconocimiento del habla, la toma de decisiones, la traducción de idiomas y el reconocimiento de patrones [30].

El término "inteligencia artificial" fue definido inicialmente por John McCarthy en 1956 durante la Conferencia de Dartmouth, que es conocido históricamente como el punto de partida del campo de la inteligencia artificial.

La inteligencia artificial ha revolucionado múltiples industrias al proporcionar soluciones eficientes y precisas a problemas difíciles. Permite automatizar procesos, mejorar la toma de decisiones, personalizar experiencias de usuario y detectar patrones en enormes volúmenes de datos. Esto ha provocado mejoras significativas en productividad, innovación y calidad de vida y se aplica en diferentes áreas como la salud, finanzas, transporte y entretenimiento.

Las técnicas de inteligencia artificial se dividen en diversas áreas entre las que sobresalen el *machine learning*, *deep learning*, procesamiento del lenguaje natural, visión por computadora y sistemas de recomendación [32]. A continuación, se definen las áreas que se utilizan en este proyecto.

3.3. *Machine learning*

El aprendizaje automático (*machine learning*) es considerada una subdisciplina de la inteligencia artificial que se dedica al desarrollo de algoritmos que permiten a las computadoras tener la capacidad de aprender a partir de datos y realizar predicciones o tomar decisiones sin haber sido programadas para llevar esas tareas específicamente [12]. A continuación, se detallan los tipos que existen:

- **Aprendizaje supervisado:** aquí los algoritmos se entrenan con un conjunto de datos etiquetados, lo que quiere decir que cada ejemplo de entrenamiento está relacionado con una etiqueta. En esta área destacan la regresión lineal y logística, los árboles de decisión y bosques aleatorios y por último, las máquinas de vectores soporte.
- **Aprendizaje no supervisado:** en esta sección, los algoritmos trabajan con datos que no están etiquetados y el objetivo es detectar patrones o estructuras ocultas en los datos. En esta área destaca el algoritmo de Kmeans.
- **Aprendizaje por refuerzo:** finalmente, aquí los agentes aprenden a tomar decisiones al realizar una interacción con su entorno y reciben recompensas o castigos dependiendo de sus acciones. En esta área destaca el algoritmo de *Q-learning*.

En este proyecto únicamente se van a utilizar algoritmos de aprendizaje supervisado ya que el conjunto de datos está etiquetado. Por otro lado, se afronta un problema de clasificación porque a los datos se les asigna una de varias clases posibles, es decir, el objetivo es predecir la categoría o etiqueta correcta para cada dato entre múltiples clases predefinidas como puede ser el ganador del partido o el número de goles.

Los algoritmos de aprendizaje supervisado asociados a problemas de clasificación que se utilizan para entrenar los modelos en este proyecto con los datos obtenidos son los siguientes [31]:

- **Árboles de decisión:** se encargan de dividir iterativamente el conjunto de datos en subconjuntos basados en las características más determinantes e importantes, estableciendo una estructura de árbol donde las hojas representan los resultados de las decisiones que se han tomado y los nodos representan los atributos.
- **Máquinas de vectores de soporte (SVM):** es un algoritmo supervisado que encuentra el hiperplano óptimo que permite separar las clases en el espacio de características. Para problemas no lineales, SVM puede utilizar trucos de kernel para así proyectar los datos a un espacio de una dimensión mayor donde puede que las clases sean linealmente separables.
- **k-vecinos más cercanos (k-NN):** k-NN es un algoritmo supervisado que permite predecir el valor de una nueva instancia en base a los k ejemplos más cercanos en

el espacio de características. Este algoritmo no contiene una fase de entrenamiento explícita como tal, lo que lo hace simple y eficaz para conjuntos de datos pequeños.

- ***Gradient boosting machines* (GBM):** es un algoritmo supervisado que permite crear modelos predictivos a través de la construcción secuencial de árboles de decisión, donde cada árbol creado se encarga de corregir los errores que ha cometido el anterior. Los ejemplos más utilizados son XGBoost y LightGBM, que son bastante eficaces y permiten manejar grandes conjuntos de datos con alta dimensionalidad.
- **Bosques aleatorios:** es un algoritmo supervisado que genera múltiples árboles de decisión entrenados en diversos subconjuntos del conjunto de datos y características. La predicción final la realiza mediante un proceso de agregación donde se utiliza la votación para clasificación, lo que ayuda a incrementar la precisión y disminuye el sobreajuste. Destaca por ser robusto y eficaz para manejar conjuntos de datos grandes y complejos.
- **Gaussian Naive Bayes:** es un algoritmo de clasificación supervisada que se basa en el teorema de Bayes, que asume la independencia entre las características. A pesar de esta suposición de cierta manera simplificadora, es bastante eficaz y computacionalmente eficiente para problemas de clasificación de datos categóricos.

Se pueden utilizar diferentes métricas para evaluar los modelos entrenados con estos algoritmos sobre el conjunto de datos. Para este proyecto, a continuación, se detallan las principales métricas que existen [25]:

- **Exactitud (*Accuracy*):** se define como la proporción de predicciones correctas entre el total de predicciones realizadas. Se calcula como (TP + TN) / (TP + TN + FP + FN), donde TP son los verdaderos positivos, TN son los verdaderos negativos, FP son los falsos positivos y FN son los falsos negativos.
- **Precisión (*Precision*):** se encarga de medir la proporción de verdaderos positivos entre las predicciones positivas. Se calcula como TP / (TP + FP). Determina la exactitud del clasificador al identificar verdaderos positivos, siendo fundamental en conjuntos de datos donde los falsos positivos tienen un alto valor.
- **Exhaustividad (*Recall*):** es la proporción de verdaderos positivos detectados correctamente entre todos los casos reales positivos considerados. Se calcula como TP / (TP + FN). Destaca su importancia en situaciones donde es crítico capturar todos los verdaderos positivos.
- **Puntuación F1 (*F1 Score*):** es la media armónica entre la precisión y la exhaustividad, estableciendo un balance o equilibrio entre ambas métricas. Se calcula como 2 * (Precisión * Exhaustividad) / (Precisión + Exhaustividad). Es adecuada cuando se requiere un equilibrio entre precisión y exhaustividad.

- **Matriz de confusion (*Confusion matrix*):** es una tabla que ayuda a visualizar el rendimiento de un modelo de clasificación. Tiene cuatro cuadrantes: TP, TN, FP, y FN, que representan las verdaderas y falsas predicciones para las clases positivas y negativas que existen. Describe una visión detallada de manera gráfica de cómo el modelo clasifica cada clase, ayudando a realizar el análisis de errores específicos de una clase.

En este caso, para este proyecto para evaluar la calidad de los modelos entrenados se va a utilizar la exactitud de manera inicial, pero en caso de que aparezcan modelos con valores similares en esta métrica, se utilizarán el resto de las métricas para realizar el desempate y evaluar qué modelo es mejor.

3.4. *Deep learning*

El aprendizaje profundo (*Deep learning*) es una subdisciplina del aprendizaje automático que se dedica al uso de redes neuronales artificiales con varias capas profundas para modelar y comprender patrones complejos en los datos que se quieran analizar. Se ha popularizado en los últimos años debido a su capacidad para superar a otros algoritmos en múltiples tareas como el reconocimiento de imágenes, procesamiento del lenguaje natural y otros campos [34]. A continuación se detallan los componentes de la arquitectura de una red neuronal [13]:

- **Neuronas artificiales:** simulan el funcionamiento de las neuronas biológicas que tienen los humanos. Cada neurona recibe varias entradas, las procesa mediante la denominada función de activación y produce una salida.
- **Capas:** las redes neuronales están compuestas por capas de neuronas. Las capas comunes incluyen la capa de entrada, capas ocultas y la capa de salida.
- **Funciones de activación:** introducen no linealidad en la red, dando la capacidad de que se modelen relaciones complejas. Ejemplos incluyen "ReLU" (*Rectified Linear Unit*), "Sigmoide" y "Tanh".

A continuación se detallan los principales tipos de redes neuronales que existen [3]:

- **Redes neuronales artificiales (ANN):** donde destaca el perceptrón multicapa, que está compuesto por una capa de entrada, una o más capas ocultas y una capa de salida. Se entrena aplicando un proceso de retropropagación (*backpropagation*) y optimización. La *backpropagation* es un método de entrenamiento que se encarga de ajustar los pesos de las conexiones neuronales reduciendo el error entre las predicciones y los valores reales del conjunto de datos. Su tarea es calcular el gradiente del error con respecto a cada peso mediante la aplicación la regla de la cadena, propagando el error desde la salida hacia la entrada.

- **Redes neuronales convolucionales (CNN):** su arquitectura se divide en capas convolucionales que aplican diversos filtros para extraer características locales de las imágenes, como pueden ser los bordes y texturas. Además incorpora capas de *pooling* que reducen la dimensionalidad de las características que se han extraído, garantizando que se mantiene la información importante y disminuyendo el coste computacional. Finalmente, incorpora capas completamente conectadas que permiten conectar todas las neuronas de una capa a todas las neuronas de la siguiente capa, como en una ANN tradicional.
- **Redes neuronales recurrentes (RNN):** su arquitectura permite que la red contenga una memoria interna y sea capaz de procesar secuencias de datos al utilizar su salida como entrada en el siguiente paso temporal que va a realizar. En este caso destaca la LSTM que es una variante de la RNN que fue diseñada para controlar dependencias a largo plazo mediante la incorporación de celdas de memoria y diferentes puertas como las puertas de entrada, olvido y salida.

De entre estos tres tipos, para este proyecto, por la naturaleza de los datos, solo se van a entrenar redes neuronales artificiales. Por otro lado, las principales métricas de evaluación para las redes neuronales coinciden con las de los modelos entrenados con algoritmos de *machine learning*.

Sin embargo, para las redes neuronales se utilizan algoritmos de optimización para ajustar los pesos de la red con el objetivo de minimizar la función de pérdida, incrementando así la calidad del modelo. El optimizador ayuda a la red a aprender patrones en los datos realizando iteraciones y ajustes de manera incremental [15]. Los principales algoritmos de optimización para redes neuronales en problemas de clasificación son los siguientes:

- **Descenso de gradiente estocástico (SGD):** actualiza los pesos de la red neuronal usando el gradiente del error calculado en cada mini-lote de datos, realizando una actualización de manera más frecuente y rápida pero que por el contrario, puede ser ruidosa. Sus ventajas son que es simple y eficiente para grandes conjuntos de datos.
- **Adam (*Adaptive Moment Estimation*):** junta las ventajas de "AdaGrad" y "RMSProp", modificando las tasas de aprendizaje para cada parámetro en base a estimaciones de primer y segundo momento del gradiente. Sus ventajas son que es robusto y eficaz para problemas con grandes volúmenes de datos y alta dimensionalidad.
- **RMSProp (*Root Mean Square Propagation*):** ajusta la tasa de aprendizaje para cada parámetro de forma adaptativa, dividiendo el gradiente por la media móvil de magnitudes recientes de este gradiente que se está analizando. Es útil para manejar la tasa de aprendizaje en problemas donde los gradientes son cambiantes y varían su valor.

Por otro lado, la función de pérdida de una red neuronal se utiliza para contar la discrepancia entre las predicciones del modelo y los valores reales. Se utiliza como guía

para el ajuste de los pesos de la red durante el entrenamiento, colaborando de esta forma a mejorar la precisión del modelo [14]. Las principales funciones de pérdida para problemas de clasificación que existen son:

- **Entropía cruzada (*Cross-Entropy Loss*):** mide la discrepancia entre las distribuciones de probabilidad predicha y la real, penalizando considerablemente las predicciones que se hayan realizado de manera incorrecta. Esta función de pérdida es ampliamente utilizada para problemas de clasificación binaria y multiclase mayoritariamente, permitiendo ajustar las probabilidades predichas a las verdaderas.
- ***Categorical Cross-Entropy:*** es una forma específica de la entropía cruzada que se utiliza en clasificación multiclase. Se calcula utilizando la probabilidad predicha para la clase verdadera y es habitual aplicarla en tareas de clasificación que afecten a imágenes y texto.
- ***Binary Cross-Entropy:*** similar a la entropía cruzada, pero específica solo para problemas de clasificación binaria. Calcula la pérdida como la media de las pérdidas individuales para cada clase binaria que se encuentre, penalizando las predicciones alejadas de las etiquetas binarias que realmente son verdaderas.
- ***Sparse Categorical Cross-Entropy:*** es una variante de la entropía cruzada categórica que se usa cuando las etiquetas están codificadas como enteros en lugar de vectores *one-hot*. Se utiliza para tareas de clasificación multiclase con un elevado número de clases.

Otro concepto importante de las redes neuronales son las épocas, que en una red neuronal determinan el número de veces que el algoritmo de entrenamiento procesa el conjunto completo de datos de entrenamiento. Cada época permite que la red ajuste sus pesos iterativamente para incrementar su rendimiento. Más épocas generalmente conducen a un mejor ajuste del modelo, aunque demasiadas pueden llevar al sobreajuste [4].

Relacionado con las épocas, se define también el tamaño de lote (*batch size*), que en una red neuronal se refiere al número de muestras de entrenamiento que son procesadas antes de actualizar los pesos del modelo. Establece la frecuencia con la que se ajustan los parámetros durante el entrenamiento. Un tamaño de lote más grande puede acelerar el entrenamiento de forma que se realice más rápido, pero consume más recursos, mientras que un tamaño de lote menor puede realizar actualizaciones más precisas pero considerablemente más lentas.

Los *callbacks* en una red neuronal son funciones personalizables que se activan en momentos específicos durante el entrenamiento, habitualmente como al terminar cada época o cuando se alcanza cierta métrica. Permiten realizar acciones como guardar el modelo, reajustar la tasa de aprendizaje o parar el entrenamiento de forma temprana según cuando se cumplan ciertas condiciones. Los *callbacks* se utilizan para monitorear y mejorar el rendimiento del modelo durante el entrenamiento y pueden ayudar a mejorar la calidad de los modelos [1].

Ademas de todos aspectos, en la estructura de una red neuronal se pueden modificar los siguientes parámetros [37]:

- ***Dropout:*** es una técnica de regularización que desactiva aleatoriamente un porcentaje de neuronas durante el entrenamiento para evitar el sobreajuste, aumentando la capacidad de generalización del modelo.
- ***Batch Normalization:*** normaliza las activaciones de una capa antes de pasar a la siguiente capa, estabilizando y acelerando el proceso de entrenamiento al disminuir el cambio de variables internas.
- ***Bias Initializer y Regularizer:*** *bias initializer* establece cómo se inicializan los sesgos en las neuronas, mientras que el *bias regularizer* aplica una penalización para evitar el sobreajuste, haciendo que se mantengan los sesgos en valores razonables durante el entrenamiento.
- ***Kernel Initializer y Regularizer:*** *kernel initializer* establece cómo se inicializan los pesos de las neuronas y el *kernel regularizer* aplica una penalización a los pesos para evitar el sobreajuste, permitiendo mantener los pesos del modelo bajo control.

3.5. Validación cruzada y división de los datos

Por otro lado, para el entrenamiento de los modelos se aplicará la técnica de validación cruzada, que es un método de validación que divide el conjunto de datos en k subconjuntos y se encarga de entrenar el modelo k veces, cada vez con un subconjunto diferente cuyo objetivo es evaluar la capacidad del modelo para generalizar a datos no vistos [7].

Al entrenar los modelos, los datos se separan en un conjunto de entrenamiento y de prueba. El conjunto de entrenamiento con un 80 % de los datos se utiliza para ajustar los parámetros del modelo. Finalmente, el conjunto de prueba con un 20 % de los datos se utiliza para evaluar la capacidad de generalización del modelo a datos no vistos. Esta división permite asegurar que el modelo no solo se comporte bien con los datos conocidos sino que también sea efectivo con datos nuevos.

4: Técnicas y herramientas

4.1. Introducción

En este capítulo se detalla qué técnica se utiliza para la extracción de los datos y qué tecnologías se utilizan para el desarrollo del proyecto.

4.2. Obtención de los datos

Para la obtención de los datos, se crean varios *scripts* en Python que extraigan los datos de la página de Resultados De Fútbol [33] mediante *scraping*.

El *scraping* [19] es una técnica utilizada para extraer automáticamente información de sitios web de forma automatizada. Consiste en el análisis y la recopilación de datos de páginas web. Estos programas acceden a la página web de la que se desean obtener los datos, identifican los componentes clave dentro del código HTML y extraen su información para su posterior procesamiento o análisis. El *scraping* es una herramienta útil para obtener datos en gran volumen de manera rápida y eficiente y es aplicada en variedad de aplicaciones. En este proyecto se ha utilizado esta técnica para obtener los datos ya que no se ha podido encontrar ningún conjunto de datos que recoja información sobre los datos históricos de los partidos en las ligas seleccionadas. Además, mediante el *scraping*, se puede fácilmente ir incorporando a los datos utilizados para crear los modelos los nuevos datos asociados a los últimos partidos jugados.

4.3. Tecnologías utilizadas

A continuación, se detallan las tecnologías base que se utilizan en el proyecto:

- **Python:** es un lenguaje de programación versátil y de alto nivel que tiene una enorme popularidad en diversos campos, destacando en la ciencia de datos. Este lenguaje incorpora diferentes bibliotecas que permiten realizar diferentes tareas lo

que le convierte en uno de los lenguajes con más funcionalidades diferentes [29]. A continuación se detallan las principales bibliotecas de Python que se utilizan en este proyecto:

- **BeautifulSoup:** sirve para realizar tareas de *scraping*. Esta biblioteca permite analizar y extraer datos de páginas web de manera sencilla y eficiente, ayudando al programador a realizar la manipulación de la estructura HTML de los sitios web para obtener la información que se desee sobre el sitio. Con BeautifulSoup, se pueden crear *scripts* que naveguen por el contenido de una página web, identifiquen los elementos deseados y que permitan extraer datos de manera automatizada [6].
- **Scikit-learn:** Python es mundialmente utilizado en el campo de la inteligencia artificial y el *machine learning* por bibliotecas como scikit-learn. Esta es una biblioteca que ofrece una diversa gama de herramientas para la creación de algoritmos de *machine learning* como se pretende en este proyecto. Con scikit-learn, se pueden crear y entrenar modelos de *machine learning* de forma eficiente, utilizando algoritmos ya definidos y técnicas avanzadas de análisis de datos [22].
- **Pandas:** facilita la manipulación y el análisis de datos estructurados mediante la introducción de los *dataframes*. Estos son estructuras de datos bidimensionales que tienen la capacidad de almacenar y manipular datos de manera eficiente, de manera similar a una tabla de base de datos o una hoja de cálculo. Con pandas, se pueden cargar datos desde multitud de fuentes, realizar operaciones de limpieza y transformación de datos y realizar análisis estadísticos y exploratorios de manera rápida. Esto hace que pandas sea una herramienta indispensable para el almacenamiento y la manipulación de datos en proyectos de ciencia de datos y análisis de datos en Python como es en este caso [9].

- **Keras:** es una API de alto nivel para la construcción, entrenamiento y evaluación de modelos de redes neuronales mediante Python. Destaca por su facilidad de uso y su enfoque en la creación rápida y sencilla de modelos de aprendizaje profundo facilitando el trabajo a los usuarios. Ofrece una sintaxis simple y fácil de entender y una abstracción de alto nivel que permite crear modelos complejos de manera rápida sin apenas esfuerzo, lo que lo convierte en una herramienta excelente que brinda flexibilidad y potente para trabajar en una amplia gama de proyectos de inteligencia artificial y aprendizaje profundo como es lo que se pretende en este caso. Esta tecnología se utiliza en este proyecto para crear modelos de redes neuronales que pueden tener un buen rendimiento sobre el conjunto de datos proporcionado [8].

- **Tensorflow:** es una biblioteca de aprendizaje automático de código abierto que ha sido desarrollada por Google que proporciona una plataforma flexible y escalable para construir, entrenar y desplegar modelos de aprendizaje profundo con diversas características. Esta tecnología destaca por su capacidad para trabajar con grandes volúmenes de datos y su eficiencia en la ejecución en variedad de plataformas.

TensorFlow ofrece una extensa gama de herramientas y funcionalidades, incluyendo la construcción de redes neuronales convolucionales, recurrentes y generativas, así como la experimentación con técnicas avanzadas. Por lo tanto, TensorFlow es una opción popular y extensamente utilizada para proyectos de inteligencia artificial y aprendizaje automático en diversas áreas. Esta tecnología se utiliza en este proyecto para la creación de modelos más avanzados que pueden tener un rendimiento elevado sobre los datos proporcionados [17].

- **Google Colaboratory:** es una plataforma gratuita de Jupyter Notebook que permite escribir y ejecutar código Python en el navegador. Está esencialmente diseñada para la enseñanza y la investigación en *machine learning*, ya que proporciona acceso a grandes recursos de computación en la nube, incluyendo GPUs y memoria RAM. Colab también ayuda a la colaboración en tiempo real, permitiendo compartir y editar *notebooks* entre diferentes usuarios. Además, se integra perfectamente con Google Drive para el almacenamiento y la gestión de archivos [10].

5: Aspectos relevantes del desarrollo del proyecto

5.1. Introducción

En este capítulo se recogen los aspectos más interesantes del desarrollo del proyecto, donde se documenta desde la metodología aplicada para el desarrollo del proyecto, describiendo los pasos a seguir y el alcance que se espera que tenga el proyecto. Por otro lado, también se detalla la planificación del proyecto, cuál es el modelo de los datos, qué operaciones de transformación y limpieza han sido necesarias y finalmente se describe cómo se ha implementado el proyecto describiendo los archivos y carpetas que forman parte del proyecto.

5.2. Metodología

La metodología de este proyecto tiene como base un enfoque que comprende varias etapas clave. En primer lugar, se realizará una revisión de diferentes artículos sobre técnicas de inteligencia artificial aplicadas al análisis de datos deportivos, centrándose especialmente en la predicción del rendimiento de equipos de fútbol basándose en la rotación de los jugadores. Esta revisión ayudará a identificar las mejores prácticas y los enfoques que pueden ser más relevantes para el desarrollo del proyecto.

Posteriormente, se llevará a cabo la recopilación y preparación de datos, donde se recogerán conjuntos de datos históricos que abarquen información relevante sobre la rotación de jugadores y el rendimiento deportivo de equipos de fútbol en las ligas seleccionadas. Esta etapa incluye la limpieza de datos, la preparación de los datos para su análisis posterior y un breve análisis sobre ellos para detectar patrones.

Una vez preparados los datos, se realizará la implementación y evaluación de modelos de inteligencia artificial. Se probarán los diferentes algoritmos comentados en la parte teórica y diferentes redes neuronales utilizando los parámetros también comentados. Los modelos se entrenarán y ajustarán utilizando los datos que se han obtenido previamente y

se evaluará su rendimiento utilizando las métricas enumeradas en la parte teórica. Esta fase permitirá detectar los modelos más eficaces y precisos para predecir el rendimiento deportivo basado en la rotación de jugadores.

Finalmente, se tratará de optimizar al máximo el rendimiento de los mejores modelos seleccionados buscando los parámetros que hagan que estos modelos tengan mejores valores en las métricas seleccionadas y de esta manera sean más precisos y puedan ayudar a los directivos en la toma de decisiones sobre las estrategias de gestión de jugadores.

5.3. Alcance

El alcance de este proyecto abarca la evaluación y aplicación de diversas técnicas de inteligencia artificial para predecir el rendimiento de equipos de fútbol basándose en la rotación de jugadores. En primer lugar, después de definir la metodología, se seleccionarán las técnicas más correctas para el análisis de datos relacionados con la rotación de jugadores y el rendimiento deportivo. Este apartado incluirá la recopilación, preprocesamiento y análisis de los datos de las ligas, equipos y jugadores de fútbol seleccionados. Para la parte del análisis de los datos, se realizan unos pequeños programas que analicen los datos obtenidos mediante mapas de calor para poder detectar patrones. Al detectar estos patrones se pretende justificar si las estrategias de rotación aplicadas por los equipos mejoran el rendimiento o no.

Las ligas sobre las que se obtendrán y utilizarán los datos serán LaLiga EA Sports (primera división española), Premier League (primera división inglesa) y Bundesliga (primera división alemana) desde la temporada 2018/2019 hasta la temporada 2023/2024, ambas incluidas. Esta variedad en la elección de ligas y temporadas permite evaluar si existen diferencias significativas entre las ligas de los diferentes países o entre las temporadas. Otra ventaja es que al utilizar datos de diferentes ligas y temporadas, probablemente los modelos creados sean más robutos y tengan mayor capacidad de generalización.

Además, el alcance del proyecto se pretende que también implique la implementación y ajuste de modelos de inteligencia artificial para la predicción del rendimiento deportivo en función de la rotación de jugadores. Para ello, se explorarán diversas técnicas de inteligencia artificial y *machine learning*, como redes neuronales, árboles de decisión y métodos de aprendizaje automático supervisado, con el objetivo de detectar aquellas que mejor se adapten a las características de este problema. Sobre cada una de ellas, se realizará una optimización de parámetros para mejorar todo lo posible su precisión.

Finalmente, se realizará una evaluación de los modelos desarrollados, utilizando métricas de rendimiento para definir su calidad en la predicción del rendimiento de los equipos. Después de esto, se seleccionarán los mejores modelos.

Se pretenden obtener tres modelos, uno entrenado para predecir el ganador del partido, otro para predecir el número de goles que anotará el equipo local y otro para predecir el número de goles que anotará el equipo visitante.

Semana	Fecha de inicio	Fecha de fin	Carga de trabajo	Sección
1	29/04/2024	05/05/2024	40 horas	I+D+i
2	06/05/2024	12/05/2024	40 horas	I+D+i
3	13/05/2024	19/05/2024	40 horas	I+D+i
4	20/05/2024	26/05/2024	30 horas	I+D+i
5	27/05/2024	02/06/2024	24 horas	I+D+i
6	03/06/2024	09/06/2024	16 horas	I+D+i
7	10/06/2024	16/06/2024	60 horas	TFM
8	17/06/2024	23/06/2024	60 horas	TFM
9	24/06/2024	28/06/2024	30 horas	TFM

Tabla 5.1: Planificación de las semanas.

Es cierto que se pretenden obtener los modelos lo más precisos posibles y que realicen las mejores predicciones, pero sin embargo, el objetivo principal del proyecto es determinar qué estrategias y qué acciones sobre la gestión de los jugadores contribuyen al mejor rendimiento de los equipos de fútbol. Por lo tanto, no se debe desviar el foco y plantear que los modelos sean capaces de realizar predicciones exactas sobre los partidos, ya que por un lado, este no es el objetivo y por otro lado, esto es algo bastante complejo debido a la incertidumbre que hay en cada partido de fútbol y la aleatoriedad que rodea cada evento de este tipo.

Además de todos estos aspectos comentados, se documentarán y analizarán todas las tareas realizadas en el proyecto, con el objetivo de ofrecer recomendaciones para la gestión de la rotación de jugadores en equipos de fútbol, así como posibles áreas de mejora y futuras investigaciones para este proyecto.

5.4. Plan de proyecto

El proyecto comienza las primeras semanas durante la estancia en un GIR para la asignatura de I+D+i donde se realiza una parte de investigación y se desarrolla el núcleo del proyecto. Para finalizar, el proyecto continua como Trabajo de Fin de Máster, donde se sigue profundizando y expandiendo el trabajo realizado previamente. La duración de cada una de estas secciones es 190 horas y 150 horas respectivamente, sumando en total 340 horas. La fecha de inicio del proyecto es el 29 de abril de 2024 y la fecha límite de finalización es el 28 de junio de 2024.

La Tabla 5.1 muestra la planificación de las semanas durante las que se desarrolla el proyecto.

En este proyecto se ha aplicado *Scrum* desde el inicio. *Scrum* es un marco ágil de gestión de proyectos que se aplica para desarrollar, entregar y mantener productos complejos. Está basado en la colaboración, la flexibilidad y la entrega incremental de productos. Scrum se encarga de dividir el trabajo en ciclos cortos llamados *sprints*, que suelen durar entre

dos y cuatro semanas. Cada *sprint* comienza con una reunión de planificación donde se define el trabajo que se realizará y termina con una revisión y retrospectiva para analizar el progreso y mejorar los procesos realizados. Los roles clave en Scrum incluyen el *Product Owner*, que define y prioriza el trabajo que se debe realizar en el proyecto, el *Scrum Master* que facilita el proceso y se encarga de eliminar los impedimentos que puedan aparecer, y el equipo de desarrollo, que es responsable de desarrollar y entregar el producto de forma incrementable [24].

En proyectos de ciencia de datos, como es el que se pretende realizar en este caso, *Scrum* se aplica para gestionar el ciclo de vida del desarrollo de los modelos y el análisis. Durante cada *sprint*, el equipo puede enfrentar tareas como la recolección y limpieza de datos, el desarrollo y entrenamiento de modelos y la validación de resultados. El *Product Owner* en este contexto puede ser un especialista en datos que determina las prioridades basándose según los objetivos del proyecto que se han planteado conseguir, mientras que el *Scrum Master* asegura que el equipo pueda trabajar de manera eficiente eliminando todos los obstáculos y facilitando la comunicación entre el equipo. Las revisiones y retrospectivas permiten ajustar rápidamente las estrategias basadas en los resultados que se han obtenido, garantizando que el proyecto se mantenga siempre alineado con las metas planteadas y pueda adaptarse a nuevas necesidades y requerimientos [26].

Para este proyecto se ha aplicado *Scrum* pero con diversas adaptaciones para ajustarlo a la naturaleza del proyecto. El primer lugar, la duración de los *sprints* ha sido de una semana según la planificación que se puede apreciar en la Tabla 5.1. Por otra parte, el alumno que se ha encargado de desarrollar el proyecto ha cumplido los roles de *Product Owner*, *Scrum Master* y equipo de desarrollo. Finalmente, se han realizado reuniones de planificación antes de comenzar los *sprints* donde se definía el trabajo a realizar y reuniones de revisión al finalizar los *sprints* para revisar los resultados obtenidos y sugerir mejoras.

5.5. Modelo de los datos

Los datos obtenidos se asocian a diferentes entidades que están relacionadas entre sí y abarcan multitud de campos, por ello, es importante estructurarlos de la manera correcta para que puedan ser utilizados adecuadamente en el entrenamiento de los modelos. En la figura 5.1 se puede apreciar el modelo de los datos y como se han almacenado de forma estructurada después de extraerlos mediante *scraping*.

Como se puede ver, en el diagrama los atributos de la entidad "IndicadoresEquipoPrepartidoModelo" no están incorporados ya que contiene 174 atributos. Más adelante se describen estos atributos.

A continuación, se realiza una breve descripción de cada entidad:

- **"Equipo":** recoge la información de cada equipo en una determinada liga y temporada. Cada equipo se identifica con un id único.

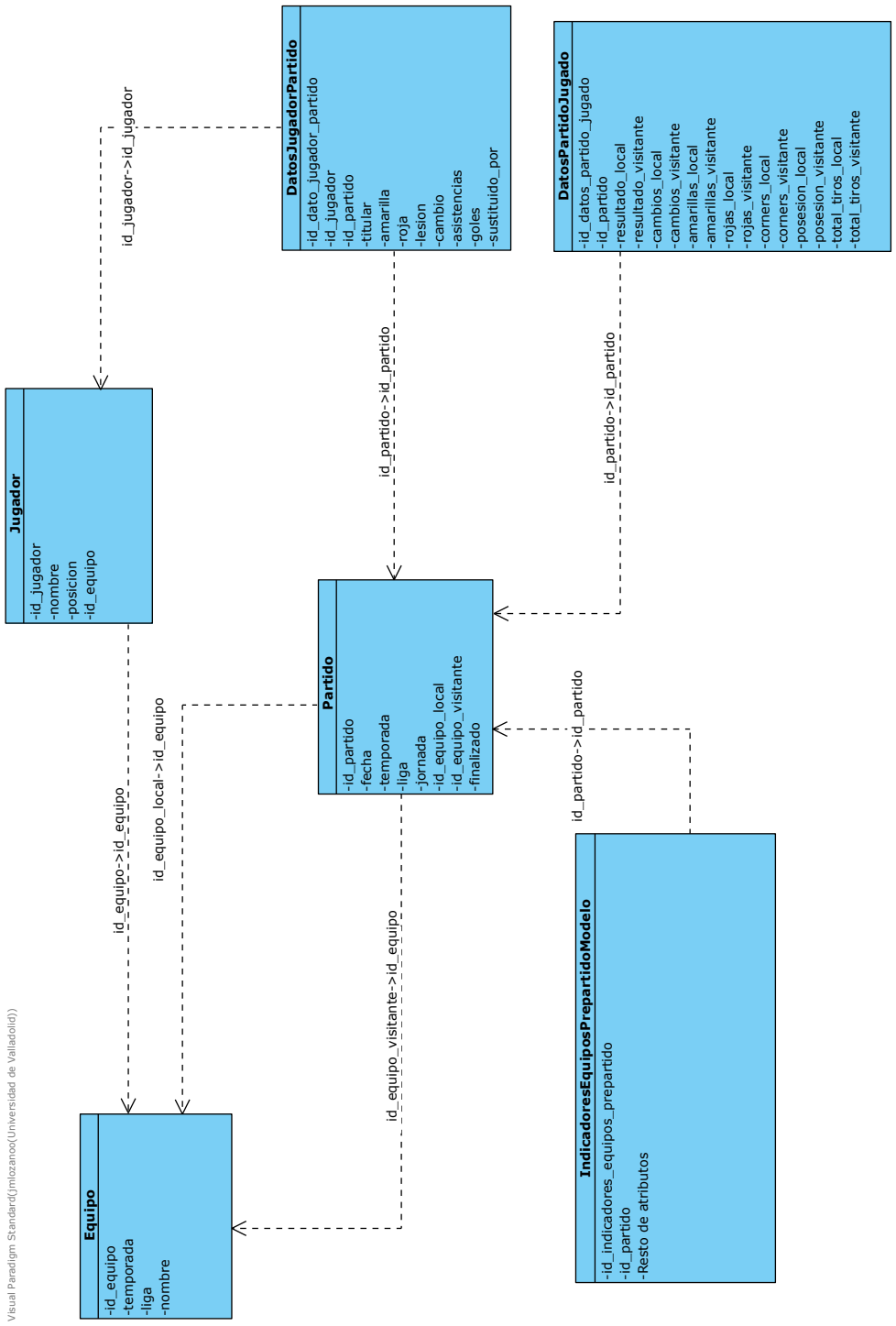


Figura 5.1: Modelo de datos.

- **"Jugador":** recoge la información de cada jugador en una determinada liga y temporada. Cada jugador se identifica con un id único y se le relaciona con el equipo en el que juega.
- **"Partido":** recoge la información de cada partido en una determinada liga y temporada. Cada partido se identifica con un id único y se le relaciona con los equipos que lo juegan.
- **"DatosJugadorPartido":** recoge la información de un jugador en un determinado partido en una determinada liga y temporada. Cada elemento se identifica con un id único y se le relaciona con el jugador y el partido al que se asocia.
- **"DatosPartidoJugado":** recoge la información de un partido jugado en una determinada liga y temporada. Cada elemento se identifica con un id único y se relaciona con el partido al que se asocia.
- **"IndicadoresEquipoPrepartidoModelo":** recoge los indicadores de los equipos que juegan un partido en una determinada liga y temporada. Cada elemento se identifica con un id único y se relaciona con el partido al que se asocia.

Los elementos de la entidad "IndicadoresEquipoPrepartidoModelo" son los datos con los que se entrenan los modelos. Los indicadores que se incluyen en esta identidad miden el desempeño previo de los equipos antes del partido. Estos indicadores pueden ser en forma de porcentaje, proporción o media y pueden tener en cuenta los partidos previos de los equipos de forma general en la temporada actual, es decir, sin distinguir si el equipo jugaba como local o visitante, o de forma específica, solo evaluando los partidos previos del equipo donde ha jugado en el mismo ámbito como lo va a hacer en el partido actual. Por ejemplo, para el equipo local de un partido, solo se tendrían en cuenta los partidos previos que ha jugado como local.

Respecto a los indicadores creados, miden diferentes valores asociados a las victorias, goles, cambios realizados, tarjetas... de los equipos. Estos indicadores definen como llegan los equipos al partido y por lo tanto son las variables explicativas. Sobre los cambios realizados, se evalúan entre otras cosas las proporciones de cada equipo de realizar cambios en unos determinados intervalos de tiempo, las proporciones de cambios según las posiciones de los jugadores afectados o las proporciones de cambios en la alineación inicial también por posiciones.

Existen tres tipos de indicadores, en forma de porcentaje, proporción o media. A continuación se detalla el cálculo de cada uno de ellos:

- **Indicadores en forma de porcentaje:** estos indicadores se calculan midiendo en qué porcentaje sobre el total de partidos jugados que se evalúen, se cumple una determinada condición. Por ejemplo, para el cálculo del porcentaje de partidos perdidos del visitante en el sitio, se calcula el porcentaje de cuántos partidos ha perdido el visitante jugando como visitante sobre cuántos partidos ha jugado el visitante como visitante en total.

- **Indicadores en forma de proporción:** estos indicadores se realizan calculando la cuenta total de un dato dividiéndolos entre el número de partidos jugados que se evalúen. Por ejemplo, para el cálculo de la proporción de puntos del local en general, se acumulan todos los puntos que haya obtenido el local jugando como local y visitante, es decir en todos sus partidos de la temporada, y se divide este valor entre el número de partidos que ha jugado el local tanto como local y visitante, es decir el total de partidos.
- **Indicadores en forma de media:** estos indicadores se calculan realizando la media sobre los datos recogidos de un determinado parámetro. Por ejemplo para la media del minuto en la que el local realiza los cambios en general, se calcula la media sobre los valores de los minutos de todos los cambios que ha realizado el local en todos sus partidos, ya sea de local o de visitante.

Además, en cada uno de estos elementos de esta entidad se incluyen las variables a predecir que son los goles de cada equipo y el ganador del partido.

A continuación, se describen todos los atributos de esta entidad "IndicadoresEquipoPrepartidoModelo" agrupándolos según del tipo que sean:

- Atributos base: atributos descriptivos de cada registro.
  - id indicadores equipo prepartido
  - id partido
  - jornada
- Atributos sobre el ganador: atributos relacionados con los ganadores de los partidos que han jugado.
  - porcentaje del local de partidos ganados en sitio
  - porcentaje del local de partidos ganados en general
  - porcentaje del local de partidos empatados en sitio
  - porcentaje del local de partidos empatados en general
  - porcentaje del local de partidos perdidos en sitio
  - porcentaje del local de partidos perdidos en general
  - porcentaje del visitante de partidos ganados en sitio
  - porcentaje del visitante de partidos ganados en general
  - porcentaje del visitante de partidos empatados en sitio
  - porcentaje del visitante de partidos empatados en general
  - porcentaje del visitante de partidos perdidos en sitio
  - porcentaje del visitante de partidos perdidos en general

  - proporción del local de puntos obtenidos en sitio
  - proporción del local de puntos obtenidos en general
  - proporción del visitante de puntos obtenidos en sitio
  - proporción del visitante de puntos obtenidos en general
- Atributos sobre la cantidad de goles: atributos relacionados con la cantidad de goles totales de los partidos que han jugado.
  - porcentaje del local de partidos con más 1,5 goles en sitio
  - porcentaje del local de partidos con más 1,5 goles en general
  - porcentaje del visitante de partidos con más 1,5 goles en sitio
  - porcentaje del visitante de partidos con más 1,5 goles en general
  - porcentaje del local de partidos con más 2,5 goles en sitio
  - porcentaje del local de partidos con más 2,5 goles en general
  - porcentaje del visitante de partidos con más 2,5 goles en sitio
  - porcentaje del visitante de partidos con más 2,5 goles en general
  - porcentaje del local de partidos con más 3,5 goles en sitio
  - porcentaje del local de partidos con más 3,5 goles en general
  - porcentaje del visitante de partidos con más 3,5 goles en sitio
  - porcentaje del visitante de partidos con más 3,5 goles en general
  - porcentaje del local de partidos con más 4,5 goles en sitio
  - porcentaje del local de partidos con más 4,5 goles en general
  - porcentaje del visitante de partidos con más 4,5 goles en sitio
  - porcentaje del visitante de partidos con más 4,5 goles en general
- Atributos sobre los goles del local: atributos relacionados con las proporciones de goles que hay en los partidos del local.
  - proporción del local de goles totales en sitio
  - proporción del local de goles totales en general
  - proporción del local de goles marcados en sitio
  - proporción del local de goles marcados en general
  - proporción del local de goles encajados en sitio
  - proporción del local de goles encajados en general
- Atributos sobre los goles del visitante: atributos relacionados con las proporciones de goles que hay en los partidos del visitante.
  - proporción del visitante de goles totales en sitio

    - proporción del visitante de goles totales en general
    - proporción del visitante de goles marcados en sitio
    - proporción del visitante de goles marcados en general
    - proporción del visitante de goles encajados en sitio
    - proporción del visitante de goles encajados en general

- Atributos sobre los goles marcados por el local: atributos relacionados con la cantidad de goles que marca el local en los partidos que juega.
    - porcentaje del local de más 0,5 goles marcados en sitio
    - porcentaje del local de más 0,5 goles marcados en general
    - porcentaje del local de más 1,5 goles marcados en sitio
    - porcentaje del local de más 1,5 goles marcados en general
    - porcentaje del local de más 2,5 goles marcados en sitio
    - porcentaje del local de más 2,5 goles marcados en general
- Atributos sobre los goles encajados por el local: atributos relacionados con la cantidad de goles que encaja el local en los partidos que juega.
    - porcentaje del local de más 0,5 goles encajados en sitio
    - porcentaje del local de más 0,5 goles encajados en general
    - porcentaje del local de más 1,5 goles encajados en sitio
    - porcentaje del local de más 1,5 goles encajados en general
    - porcentaje del local de más 2,5 goles encajados en sitio
    - porcentaje del local de más 2,5 goles encajados en general
- Atributos sobre los goles marcados por el visitante: atributos relacionados con la cantidad de goles que marca el visitante en los partidos que juega.
    - porcentaje del visitante de más 0,5 goles marcados en sitio
    - porcentaje del visitante de más 0,5 goles marcados en general
    - porcentaje del visitante de más 1,5 goles marcados en sitio
    - porcentaje del visitante de más 1,5 goles marcados en general
    - porcentaje del visitante de más 2,5 goles marcados en sitio
    - porcentaje del visitante de más 2,5 goles marcados en general
- Atributos sobre los goles encajados por el visitante: atributos relacionados con la cantidad de goles que encaja el visitante en los partidos que juega.
    - porcentaje del visitante de más 0,5 goles encajados en sitio

  - porcentaje del visitante de más 0,5 goles encajados en general
  - porcentaje del visitante de más 1,5 goles encajados en sitio
  - porcentaje del visitante de más 1,5 goles encajados en general
  - porcentaje del visitante de más 2,5 goles encajados en sitio
  - porcentaje del visitante de más 2,5 goles encajados en general
- Atributos sobre las amarillas: atributos relacionados con la proporción de amarillas que reciben los equipos en los partidos que juegan.
  - proporción del local de amarillas en sitio
  - proporción del local de amarillas en general
  - proporción del visitante de amarillas en sitio
  - proporción del visitante de amarillas en general
- Atributos sobre las rojas: atributos relacionados con la proporción de rojas que reciben los equipos en los partidos que juegan.
  - proporción del local de rojas en sitio
  - proporción del local de rojas en general
  - proporción del visitante de rojas en sitio
  - proporción del visitante de rojas en general
- Atributos sobre los cambios: atributos relacionados con la proporción de cambios que realizan los equipos en los partidos que juegan.
  - proporción del local de cambios en sitio
  - proporción del local de cambios en general
  - proporción del visitante de cambios en sitio
  - proporción del visitante de cambios en general
- Atributos sobre la posesión: atributos relacionados con la proporción de posesión que tienen los equipos en los partidos que juegan.
  - proporción del local de posesión en sitio
  - proporción del local de posesión en general
  - proporción del visitante de posesión en sitio
  - proporción del visitante de posesión en general
- Atributos sobre los tiros: atributos relacionados con la proporción de tiros que realizan los equipos en los partidos que juegan.
  - proporción del local de total tiros en sitio

  - proporción del local de total tiros en general
  - proporción del visitante de total tiros en sitio
  - proporción del visitante de total tiros en general

- Atributos sobre los córneres: atributos relacionados con la proporción de córneres que realizan y reciben los equipos en los partidos que juegan.
  - proporción del local de córneres a favor en sitio
  - proporción del local de córneres a favor en general
  - proporción del visitante de córneres a favor en sitio
  - proporción del visitante de córneres a favor en general
  - proporción del local de córneres en contra en sitio
  - proporción del local de córneres en contra en general
  - proporción del visitante de córneres en contra en sitio
  - proporción del visitante de córneres en contra en general
- Atributos sobre los cambios de lesionados, amarillas, goleadores y asistentes: atributos relacionados con la proporción de cambios de diferente naturaleza que realizan los equipos en los partidos que juegan.
  - proporción del local de cambios por jugadores lesionados en sitio
  - proporción del local de cambios por jugadores lesionados en general
  - proporción del visitante de cambios por jugadores lesionados en sitio
  - proporción del visitante de cambios por jugadores lesionados en general
  - proporción del local de cambios por jugadores con amarillas en sitio
  - proporción del local de cambios por jugadores con amarillas en general
  - proporción del visitante de cambios por jugadores con amarillas en sitio
  - proporción del visitante de cambios por jugadores con amarillas en general
  - proporción del local de cambios por jugadores goleadores en sitio
  - proporción del local de cambios por jugadores goleadores en general
  - proporción del visitante de cambios por jugadores goleadores en sitio
  - proporción del visitante de cambios por jugadores goleadores en general
  - proporción del local de cambios por jugadores asistentes en sitio
  - proporción del local de cambios por jugadores asistentes en general
  - proporción del visitante de cambios por jugadores asistentes en sitio
  - proporción del visitante de cambios por jugadores asistentes en general

- Atributos sobre la media del minuto de los cambios: atributos relacionados con la media de los minutos en la que realizan los cambios los equipos en los partidos que juegan.
    - media del local de los minutos en la que realiza los cambios en sitio
    - media del local de los minutos en la que realiza los cambios en general
    - media del visitante de los minutos en la que realiza los cambios en sitio
    - media del visitante de los minutos en la que realiza los cambios en general
- Atributos sobre los cambios de delanteros a otra posición: atributos relacionados con la proporción de cambios donde sacan un delantero por otro jugador de los equipos en los partidos que juegan.
    - proporción del local de cambios de delanteros a centrocampistas en sitio
    - proporción del local de cambios de delanteros a centrocampistas en general
    - proporción del visitante de cambios de delanteros a centrocampistas en sitio
    - proporción del visitante de cambios de delanteros a centrocampistas en general
    - proporción del local de cambios de delanteros a defensas en sitio
    - proporción del local de cambios de delanteros a defensas en general
    - proporción del visitante de cambios de delanteros a defensas en sitio
    - proporción del visitante de cambios de delanteros a defensas en general
- Atributos sobre los cambios de centrocampistas a otra posición: atributos relacionados con la proporcion de cambios donde sacan un centrocampista por otro jugador de los equipos en los partidos que juegan.
    - proporción del local de cambios de centrocampistas a delanteros en sitio
    - proporción del local de cambios de centrocampistas a delanteros en general
    - proporción del visitante de cambios de centrocampistas a delanteros en sitio
    - proporción del visitante de cambios de centrocampistas a delanteros en general
    - proporción del local de cambios de centrocampistas a defensas en sitio
    - proporción del local de cambios de centrocampistas a defensas en general
    - proporción del visitante de cambios de centrocampistas a defensas en sitio
    - proporción del visitante de cambios de centrocampistas a defensas en general
- Atributos sobre los cambios de defensas a otra posición: atributos relacionados con la proporción de cambios donde sacan un defensa por otro jugador de los equipos en los partidos que juegan.
    - proporción del local de cambios de defensas a delanteros en sitio

  - proporción del local de cambios de defensas a delanteros en general
  - proporción del visitante de cambios de defensas a delanteros en sitio
  - proporción del visitante de cambios de defensas a delanteros en general
  - proporción del local de cambios de defensas a centrocampistas en sitio
  - proporción del local de cambios de defensas a centrocampistas en general
  - proporción del visitante de cambios de defensas a centrocampistas en sitio
  - proporción del visitante de cambios de defensas a centrocampistas en general

- Atributos sobre los cambios en los minutos: atributos relacionados con la proporcion de cambios en determinados rangos de tiempo de los equipos en los partidos que juegan.

  - proporción del local de cambios en los minutos antes descanso en sitio
  - proporción del local de cambios en los minutos antes descanso en general
  - proporción del visitante de cambios en los minutos antes descanso en sitio
  - proporción del visitante de cambios en los minutos antes descanso en general
  - proporción del local de cambios en los minutos 45 a 60 en sitio
  - proporción del local de cambios en los minutos 45 a 60 en general
  - proporción del visitante de cambios en los minutos 45 a 60 en sitio
  - proporción del visitante de cambios en los minutos 45 a 60 en general
  - proporción del local de cambios en los minutos 61 a 75 en sitio
  - proporción del local de cambios en los minutos 61 a 75 en general
  - proporción del visitante de cambios en los minutos 61 a 75 en sitio
  - proporción del visitante de cambios en los minutos 61 a 75 en general
  - proporción del local de cambios en los minutos 76 a final en sitio
  - proporción del local de cambios en los minutos 76 a final en general
  - proporción del visitante de cambios en los minutos 76 a final en sitio
  - proporción del visitante de cambios en los minutos 76 a final en general

- Atributos sobre los cambios en la alineación inicial: atributos relacionados con la proporción de cambios que realizan en las alineaciones iniciales los equipos en los partidos que juegan.

  - proporción del local de cambios en la alineación de defensas en sitio
  - proporción del local de cambios en la alineación de defensas en general
  - proporción del visitante de cambios en la alineación de defensas en sitio
  - proporción del visitante de cambios en la alineación de defensas en general

  - proporción del local de cambios en la alineación de centrocampistas en sitio
  - proporción del local de cambios en la alineación de centrocampistas en general
  - proporción del visitante de cambios en la alineación de centrocampistas en sitio
  - proporción del visitante de cambios en la alineación de centrocampistas en general
  - proporción del local de cambios en la alineación de delanteros en sitio
  - proporción del local de cambios en la alineación de delanteros en general
  - proporción del visitante de cambios en la alineación de delanteros en sitio
  - proporción del visitante de cambios en la alineación de delanteros en general
- Clases a predecir: clases que se pretenden predecir en base a los anteriores atributos.
  - resultado local
  - resultado visitante
  - resultado partido

5.6. Limpieza y transformación de los datos

Las tareas de limpieza y transformación de los datos para prepararlos para que puedan ser utilizados en el entrenamiento de los modelos se describen a continuación:

- **Eliminar datos sobre sustituciones de jugadores no detectados:** se han eliminado los registros de "datosJugadoresPartidos" donde no se ha podido extraer la posición sobre el jugador sustituido. Esto ha sucedido con apenas tres jugadores en todas las ligas y temporadas evaluadas y por lo tanto el número de registros afectados es mínimo.
- **Seleccionar partidos a partir de la jornada 10:** se han filtrado los datos de los partidos dejando solamente los partidos jugados desde la jornada 10 hasta el final. Esto se ha hecho ya que los datos que se tienen en cuenta para cada partido únicamente consideran los partidos previos de los equipos que disputan ese encuentro en esa temporada y por tanto, hasta la jornada 10, no se considera que existen datos suficientes para obtener conclusiones estables sobre cómo se comporta ese equipo.
- **Eliminación de ids:** para preparar los datos para entrenar los modelos, se han eliminado tanto el id del partido asociado como el id único del dato para cada registro con los datos de los indicadores para un partido.
- **Transformacion de la clase:** específicamente, antes de entrenar las redes neuronales con los datos obtenidos, se han transformado los datos de los registros de la clase a predecir, ya sea el ganador del partido, el número de goles del local o el número de goles del visitante, aplicando *one-hot* para que las redes neuronales puedan utilizar estos datos.

Resultado partido	Cuenta	Porcentaje sobre el total
Ganador local	2186	45,35
Empate	1164	24,15
Ganador visitante	1470	30,50

Tabla 5.2: Distribución de los registros para la clase del ganador del partido.

- **Normalización de los datos:** esta es una técnica de preprocesamiento que ajusta los valores de los datos para que se encuentren en un rango común que en este caso es [0, 1]. Esto mejora la eficiencia y la precisión de los algoritmos de *machine learning* al garantizar que todas las características contribuyan equitativamente en el modelo. En este caso, es crucial para evitar que características con valores más grandes dominen el modelo ya que hay atributos que pueden tomar valores muy grandes y otros valores muy pequeños.

Después de esto, en total, agrupando los datos de las tres ligas evaluadas en las temporadas comentadas, se han obtenido los datos asociados a 4820 partidos.

5.7. Análisis de las clases a predecir

En primer lugar, en la Tabla 5.2 se puede ver la distribución de los registros para la clase del ganador del partido. Aquí se puede ver que lo más habitual es que gane el equipo local.

En la Figura 5.2 se ve de manera gráfica mediante un gráfico circular cómo se reparten las proporciones de cada resultado posible en un partido sobre el total de registros.

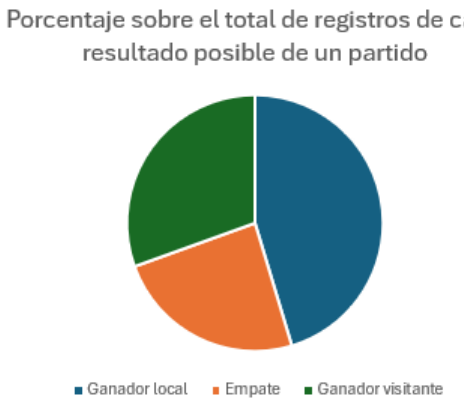


Figura 5.2: Gráfico circular con el porcentaje sobre el total de registros de cada resultado posible de un partido.

Goles marcados por el local	Cuenta	Porcentaje sobre el total
0	1057	21,93
1	1533	31,80
2	1221	25,33
3	615	12,76
4	244	5,06
5	113	2,34
6	29	0,60
7	4	0,08
8	3	0,06
9	1	0,02

Tabla 5.3: Distribución de los registros para la clase del número de goles del local en el partido.

En segundo lugar, en la Tabla 5.3 se puede ver la distribución de los registros para la clase del número de goles del local en el partido. Aquí se puede ver que lo más habitual es que marque un gol el local.

En la Figura 5.3 se ve de manera gráfica mediante un gráfico circular cómo se reparten las proporciones del número de goles anotados por el local en un partido sobre el total de registros.

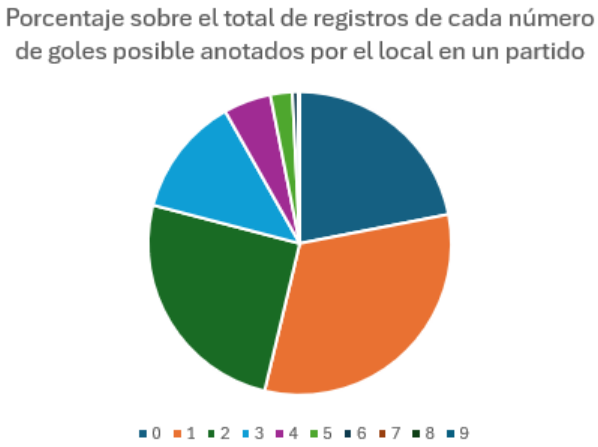


Figura 5.3: Gráfico circular con el porcentaje sobre el total de registros de cada número de goles anotados por el local posible.

En tercer lugar, en la Tabla 5.4 se puede ver la distribución de los registros para la clase del número de goles del visitante en el partido. Aquí se puede ver que lo más habitual es que marque un gol o no anote el equipo visitante.

Goles marcados por el visitante	Cuenta	Porcentaje sobre el total
0	1480	30,71
1	1674	34,73
2	995	20,64
3	421	8,73
4	181	3,76
5	49	1,02
6	18	0,37
7	1	0,02
8	0	0,00
9	1	0,02

Tabla 5.4: Distribución de los registros para la clase del número de goles del visitante en el partido.

En la Figura 5.4 se ve de manera gráfica mediante un gráfico circular cómo se reparten las proporciones del número de goles anotados por el visitante en un partido sobre el total de registros.

Figura 5.4: Gráfico circular con el porcentaje sobre el total de registros de cada número de goles anotados por el visitante posible.

5.8. Implementación

El código del proyecto se ha ejecutado en una máquina virtual proporcionada por la Escuela y en Google Colaboratory. Todo el código del proyecto se ha dividido en diferentes carpetas. A continuación se detalla la finalidad de cada una estas carpetas y sus archivos:

- **Carpeta *"scraping"*:** en esta carpeta hay diferentes archivos en Python que se encargan de realizar el *scraping* para extraer los datos de la web. Estos archivos están numerados por orden ya que cada uno se encarga de extraer unos determinados datos de una entidad. La descripción de la finalidad de cada uno de estos archivos es la siguiente:
  - **Constantes.py:** en él se definen las ligas y temporadas sobre las que se quieren extraer los datos.
  - **EjecuccionGlobal.py:** este el fichero que se debe ejecutar para obtener los datos de las ligas y temporadas definidas en el anterior fichero. Este archivo se encarga de ejecutar sucesivamente cada uno de los archivos que se comentan a continuación e ir extrayendo los datos correspondientes.
  - **1ObtencionEquiposLiga.py:** al ejecutarlo se extraen los datos de los equipos en las ligas y temporadas definidas en el fichero de constantes.
  - **2ObtencionPlantillasEquipos.py:** al ejecutarlo se extraen los datos de los jugadores en los equipos previamente obtenidos en las ligas y temporadas definidas en el fichero de constantes.
  - **3ObtencionPartidos.py:** al ejecutarlo se extraen los datos de los partidos en las ligas y temporadas definidas en el fichero de constantes.
  - **4ObtencionDatosJugadoresPartido.py:** al ejecutarlo se extraen los datos de los jugadores de los partidos en las ligas y temporadas definidas en el fichero de constantes.
  - **5ObtencionDatosPartidoJugado.py:** al ejecutarlo se extraen los datos más concretos y detallados de los partidos en las ligas y temporadas definidas en el fichero de constantes.
  - **6ObtencionIndicadoresEquiposHistorico.py:** al ejecutarlo se extraen los datos de los indicadores de los equipos en cada uno de los partidos en las ligas y temporadas definidas en el fichero de constantes.
  - **7PreparacionModelo.py:** al ejecutarlo se transforman los datos de los indicadores de cada partido en el formato adecuado para que se puedan utilizar para entrenar los modelos.
- **Carpeta "modelos":** en esta carpeta se encuentran los archivos que se encargan de entrenar los modelos. Cada archivo se asocia al entrenamiento de un modelo y el fichero con los datos con los que se entrenan los modelos se llama "datosModelo" y está en formato csv.
- **Carpeta "csv":** en esta carpeta se encuentran los datos extraídos mediante los archivos que realizan el *scraping*. Los archivos se agrupan por competición y temporada, de manera que se crean varias carpetas donde cada una contiene varios csv con los datos extraídos para cada una de las entidades previamente comentadas para esa temporada y liga.

- **Carpeta "analisis-datos":** en esta carpeta se encuentran los archivos que se encargan de realizar el análisis previo de los datos antes de entrenar los modelos para detectar patrones. En esta carpeta se pueden apreciar dos archivos ipynb que son:
  - **extraccionDatosEquipo.ipynb**: este fichero se encarga de extraer del último partido de cada equipo en una temporada, los valores de sus indicadores en general. De entrada, este fichero recibe el csv con los datos de la entidad "indicadoresEquipoHistoricoModelo", el csv con los partidos y el csv con los equipos para una liga y temporada determinada. De salida proporciona para cada equipo, su valor en cada uno de los indicadores considerados que son analizados de manera general, es decir, teniendo en cuenta todos los partidos que ha jugado el equipo en la temporada tanto como de local como de visitante.
  - **analisisAtributosEquipo.ipynb**: este fichero se encarga de calcular para cada equipo la diferencia de posiciones que existen entre la clasificación por puntos y la clasificación por cada indicador. De entrada, recibe el csv generado por el anterior fichero con los datos de cada indicador para cada equipo en una temporada y liga. De salida genera un csv donde para cada equipo y cada indicador, se recoge el número de posiciones que difiere la posición de ese equipo en la clasificación por puntos y la posición de ese equipo en la clasificación por ese indicador.

5.9. Proceso de elección de los mejores modelos

Modelos con algoritmos de *machine learning*

Con los datos obtenidos, se han entrenado diferentes algoritmos de *machine learning* de los comentados en la Sección 3.3 con los conceptos teóricos y redes neuronales variando su estructura. Para cada uno de los algoritmos comentados de *machine learning*, se calcula el valor de la exactitud para el modelo entrenado con los mejores parámetros obtenidos tras aplicar su optimización.

En este caso, para los algoritmos de *machine learning* cabe recordar que los algoritmos que se iban a evaluar eran los árboles de decisión, máquinas de vectores de soporte, k-vecinos más cercanos, *gradient boosting machines*, bosques aleatorios y Gaussian Naive Bayes. Al entrenar estos algoritmos se utiliza el conjunto de datos después de realizar las operaciones de limpieza y transformación previamente comentadas.

En la Tabla 5.5 se puede ver qué valor sobre la métrica evaluada han obtenido los diferentes modelos creados utilizando algoritmos de *machine learning* con la mejor combinación de parámetros para cada algoritmo para el problema de predecir el ganador del partido sobre el conjunto de prueba.

En este caso los dos algoritmos con mejor valor en la métrica evaluada son los bosques aleatorios y SVM, siendo ligeramente mejor el modelo que utiliza los bosques aleatorios. Para confirmar esto se han utilizado el resto de las métricas que se definieron en la parte

Algoritmo	Mejores parámetros	Exactitud (*Accuracy*) del modelo creado
Árbol de decisión	“max depth”: 3, “min samples split”: 2	0,507
SVM	“C”: 10, “gamma”: 0,001, “kernel”: rbf	0,532
K-vecinos más cercanos	“n neighbors”: 7, “p”: 2, “weights”: uniform	0,485
GBM	“learning rate”: 0,01, “max depth”: 3, “n estimators”: 50	0,494
Bosques aleatorios	“max depth”: 10, “min samples split”: 10, “n estimators”: 100	0,535
Gaussian Naive Bayes	“var smoothing”: 1e-09	0,476

Tabla 5.5: Valor de la exactitud para cada uno de los modelos entrenados utilizando algoritmos de *machine learning* para predecir el ganador del partido.

Métrica/Algoritmo	Bosques aleatorios	SVM
Precisión	0,48	0,42
Exhaustividad	0,54	0,53
Puntuación F1	0,46	0,45

Tabla 5.6: Valor del resto de las métricas sobre los dos mejores modelos para predecir el ganador del partido.

teórica para evaluar la calidad de los modelos, para de esta forma terminar de confirmar si el modelo con los bosques aleatorios es mejor y en ese caso seguir profundizando para continuar depurando sus parámetros. En la Tabla 5.6 se muestran el resto de valores de las métricas sobre los modelos entrenados con estos algoritmos.

A continuación, en la Figura 5.5 se ve la matriz de confusión para el modelo entrenado optimizando los parámetros para una SVM y en la Figura 5.6 se ve la matriz de confusión para el modelo entrenado optimizando los parámetros para un bosque aleatorio.

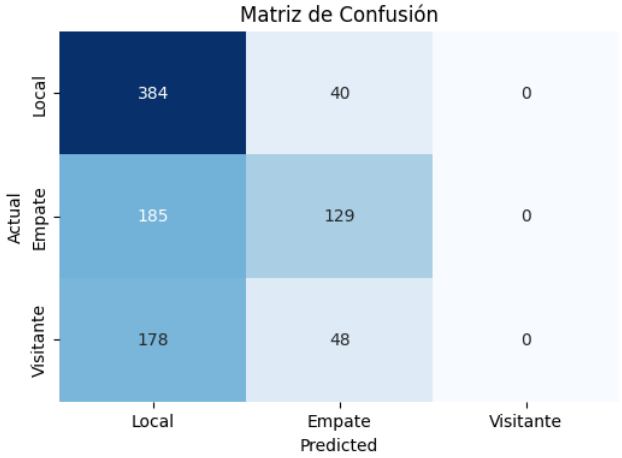


Figura 5.5: Matriz de confusión para el modelo entrenado para predecir el ganador del partido utilizando SVM.

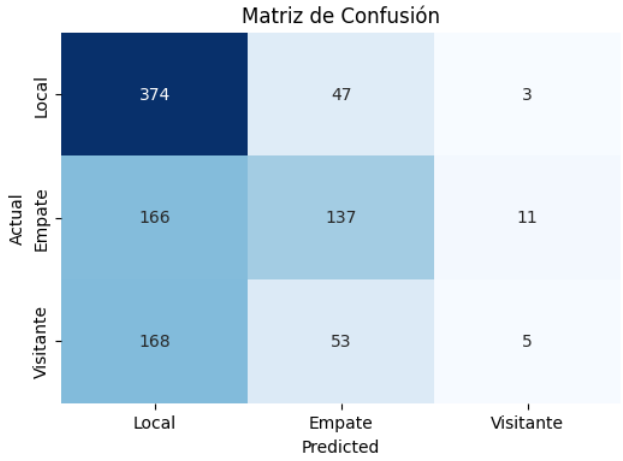


Figura 5.6: Matriz de confusión para el modelo entrenado para predecir el ganador del partido utilizando bosques aleatorios.

Por lo tanto, para el ganador del partido el modelo que proporciona mejores valores en las métricas evaluadas utiliza el algoritmo de los bosques aleatorios. Por lo tanto, se va a tratar de seguir profundizando y depurando sus parámetros para incrementar aún más su rendimiento. Para ello, se ha vuelto a realizar el proceso de entrenamiento pero en este caso, cambiando los valores posibles de los parámetros a valores más cercanos a los que se acaban de seleccionar con esta primera ejecución. Estos valores son para "max depth" 25, 30 y 35, para "min samples split" 2, 3, y 4 y para "n estimators" 190, 200 y 210. Tras el entrenamiento del modelo con validación cruzada con esta opciones posibles para los parámetros, se ha visto que el modelo finalmente creado utilizando bosques aleatorios

Algoritmo	Mejores parámetros	Exactitud (*Accuracy*) del modelo creado
Árbol de decisión	“max depth”: 3, “min samples split”: 2	0,331
SVM	“C”: 0,1, “gamma”: 0,001, “kernel”: linear	0,341
K-vecinos más cercanos	“n neighbors”: 7, “p: 1, “weights”: uniform	0,306
GBM	“learning rate”: 0,01, “max depth”: 4, “n estimators”: 50	0,315
Bosques aleatorios	“max depth”: 10, “min samples split”: 2, “n estimators”: 200	0,348
Gaussian Naive Bayes	“var smoothing”: 1e-09	0,263

Tabla 5.7: Valor de la exactitud para cada uno de los modelos entrenados utilizando algoritmos de *machine learning* para predecir los goles del local.

tiene de parámetros “max depth”: 35, “min samples split”: 3, “n estimators”: 200 y un valor sobre la exactitud de 0,542.

En la Tabla 5.7 se puede ver qué valor sobre la métrica evaluada han obtenido los diferentes modelos creados utilizando algoritmos de *machine learning* con la mejor combinación de parámetros para cada algoritmo para el problema de predecir los goles marcados por el equipo local.

En este caso los dos algoritmos con mejor valor en la métrica evaluada son los bosques aleatorios y SVM, al igual que el caso anterior, siendo ligeramente mejor el modelo que utiliza los bosques aleatorios. Para confirmar esto se han utilizado el resto de las métricas que se definieron en la parte teórica para evaluar la calidad de los modelos, para de esta forma terminar de confirmar si el modelo con los bosques aleatorios es mejor y en ese caso seguir profundizando para continuar depurando sus parámetros. En la Tabla 5.8 se muestran el resto de valores de las métricas sobre los modelos entrenados con estos algoritmos.

A continuación, en la Figura 5.7 se ve la matriz de confusión para el modelo entrenado optimizando los parámetros para una SVM y en la Figura 5.8 se ve la matriz de confusión para el modelo entrenado optimizando los parámetros para un bosque aleatorio.

Métrica/Algoritmo	Bosques aleatorios	SVM
Precisión	0,34	0,41
Exhaustividad	0,35	0,34
Puntuación F1	0,30	0,24

Tabla 5.8: Valor del resto de las métricas sobre los dos mejores modelos para predecir los goles del local.

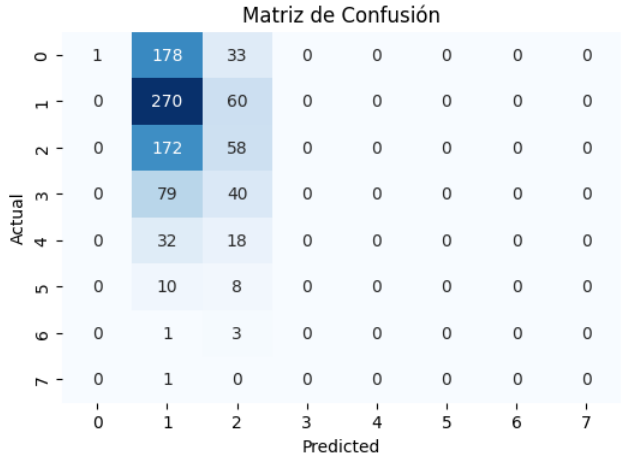


Figura 5.7: Matriz de confusión para el modelo entrenado para predecir los goles del local utilizando SVM.

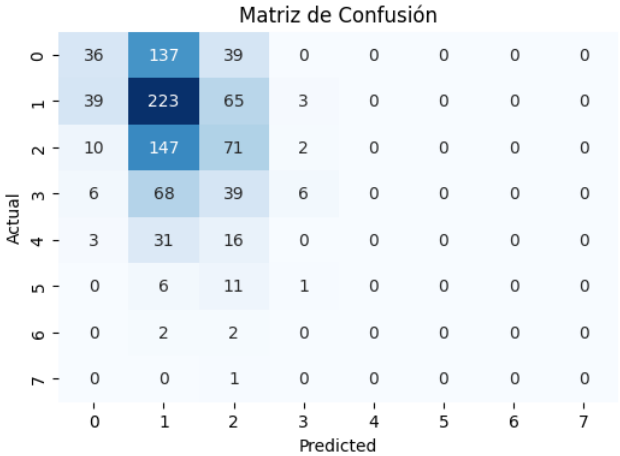


Figura 5.8: Matriz de confusión para el modelo entrenado para predecir los goles del local utilizando bosques aleatorios.

Por lo tanto, para los goles del local el modelo que proporciona mejores valores en las métricas evaluadas utiliza el algoritmo de los bosques aleatorios. Por lo tanto, se va a

Algoritmo	Mejores parámetros	Exactitud (*Accuracy*) del modelo creado
Árbol de decisión	"max depth": 3, "min samples split": 2	0,330
SVM	"C": 0,1, "gamma": 0,001, "kernel": linear	0,359
K-vecinos más cercanos	"n neighbors": 7, "p": 2, "weights": uniform	0,329
GBM	"learning rate": 0,01, "max depth": 3, "n estimators": 50	0,336
Bosques aleatorios	"max depth": 30, "min samples split": 2, "n estimators": 200	0,352
Gaussian Naive Bayes	"var smoothing": 1e-09	0,299

Tabla 5.9: Valor de la exactitud para cada uno de los modelos entrenados utilizando algoritmos de *machine learning* para predecir los goles del visitante.

tratar de seguir profundizando y depurando sus parámetros para incrementar aún más su rendimiento. Para ello, se ha vuelto a realizar el proceso de entrenamiento pero en este caso, cambiando los valores posibles de los parámetros a valores más cercanos a los que se acaban de seleccionar con esta primera ejecución. Estos valores son para "max depth" 8, 10 y 12, para "min samples split" 2, 3, y 4 y para "n estimators" 190, 200 y 210. Tras el entrenamiento del modelo con validación cruzada con esta opciones posibles para los parámetros, se ha visto que el modelo finalmente creado utilizando bosques aleatorios tiene de parámetros "max depth": 8, "min samples split": 3, "n estimators": 200 y un valor sobre la exactitud de 0,349.

En la Tabla 5.9 se puede ver qué valor sobre la métrica evaluada han obtenido los diferentes modelos creados utilizando algoritmos de *machine learning* con la mejor combinación de parámetros para cada algoritmo para el problema de predecir los goles del visitante.

En este caso los dos algoritmos con mejor valor en la métrica evaluada son los bosques aleatorios y SVM, al igual que en los casos anteriores, siendo ligeramente mejor el modelo que utiliza los bosques aleatorios. Para confirmar esto se han utilizado el resto de las métricas que se definieron en la parte teórica para evaluar la calidad de los modelos, para

Métrica/Algoritmo	Bosques aleatorios	SVM
Precisión	0,33	0,24
Exhaustividad	0,35	0,36
Puntuación F1	0,30	0,28

Tabla 5.10: Valor del resto de las métricas sobre los dos mejores modelos para predecir los goles del visitante.

de esta forma terminar de confirmar si el modelo con los bosques aleatorios es mejor y en ese caso seguir profundizando para continuar depurando sus parámetros. En la Tabla 5.10 se muestran el resto de valores de las métricas sobre los modelos entrenados con estos algoritmos.

A continuación, en la Figura 5.9 se ve la matriz de confusión para el modelo entrenado optimizando los parámetros para una SVM y en la Figura 5.10 se ve la matriz de confusión para el modelo entrenado optimizando los parámetros para un bosque aleatorio.

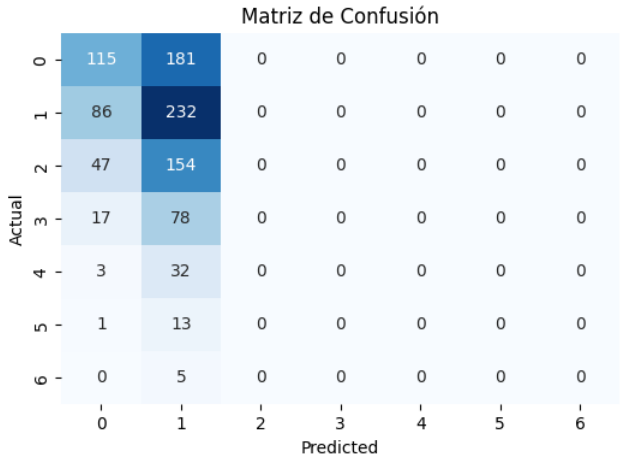


Figura 5.9: Matriz de confusión para el modelo entrenado para predecir los goles del visitante utilizando SVM.

Parámetro	Valores diferentes
Tipo de red	1,2,3,4,5,6
Tamaño de lote	16,48,80
Épocas	20,50,80
Optimizador	“Adam”, “SGD”, “RMSProp”
Callbacks	Si, No

Tabla 5.11: Parámetros a evaluar en las redes neuronales.

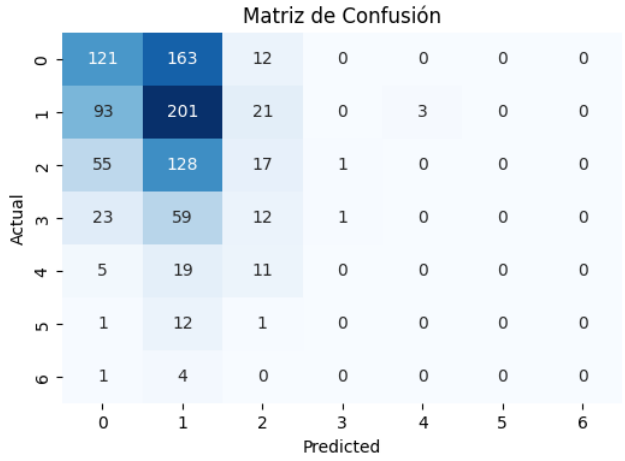


Figura 5.10: Matriz de confusión para el modelo entrenado para predecir los goles del visitante utilizando bosques aleatorios.

Por lo tanto, para los goles del visitante el modelo que proporciona mejores valores en las métricas evaluadas utiliza el algoritmo de los bosques aleatorios. Por lo tanto, se va a tratar de seguir profundizando y depurando sus parámetros para incrementar aún más su precisión. Para ello, se ha vuelto a realizar el proceso de entrenamiento pero en este caso, cambiando los valores posibles de los parámetros a valores más cercanos a los que se acaban de seleccionar con esta primera ejecución. Estos valores son para “max depth” 25, 30 y 35, para “min samples split” 2, 3, y 4 y para “n estimators” 190, 200 y 210. Tras el entrenamiento del modelo con validación cruzada con esta opciones posibles para los parámetros, se ha visto que el modelo finalmente creado utilizando bosques aleatorios tiene de parámetros “max depth”: 30, “min samples split”: 2, “n estimators”: 200 y un valor sobre la exactitud de 0,352.

Modelos con redes neuronales

Por otra parte, la otra opción que se iban a implementar eran las redes neuronales. En estos casos, para las redes neuronales, en la Tabla 5.11, se establecen los parámetros que se han optimizado sobre estas redes neuronales y los valores diferentes que podían tomar.

Previo a esta elección de parámetros, se realizó un filtrado eliminando opciones que no tenían repercusión sobre la exactitud de los modelos. Por ejemplo, se redujeron los valores del tamaño de lote y las épocas a analizar a tres valores diferentes, ya que no existía mucha diferencia entre estos valores. Por otro lado, se vió que cambiando la función de pérdida no se observaban cambios significativos en las métricas de los modelos entrenados, por lo tanto como función de pérdida se estableció por defecto *categorical crossentropy* y de esta forma también se evitaban evaluar más combinaciones de parámetros. Para el tamaño del lote se han establecido valores en un rango amplio y al igual que con las épocas, se han reducido los valores a tres opciones distintas.

A continuación, se detalla la estructura de cada uno de los tipos diferentes de red que se han evaluado:

- **Tipo de red 1:** consta de tres capas densas. La primera capa es una capa densa con 128 unidades y utilizando como función de activación "ReLU", seguida de una capa de *BatchNormalization* y una capa de *Dropout* con una tasa del 30 % para ayudar a prevenir el sobreajuste. La segunda capa es otra capa densa en este caso con 64 unidades y tambien como función de activación "ReLU", seguida por una capa de *BatchNormalization* y otra capa de *Dropout* del 30 %. La capa de salida finalmente es una capa densa con el número de unidades según el problema para el que sea y función de activación "softmax", que es la que se utiliza para problemas de clasificación multiclase.

- **Tipo de red 2:** consta de cuatro capas densas. La primera capa es una capa densa con 256 unidades y utilizando como función de activación "ReLU", seguida de una capa de *BatchNormalization* y una capa de *Dropout* con una tasa del 40 % para evitar el sobreajuste. La segunda capa es otra capa densa con 128 unidades e igualmente con la función de activación "ReLU", seguida por otra capa de *BatchNormalization* y otra capa de *Dropout* del 40 %. La tercera capa es una capa densa con 64 unidades y con función de activación "ReLU", seguida además de una capa de *BatchNormalization* y una capa de *Dropout* del 40 %. La capa de salida es una capa densa con el número de unidades según el problema para el que sea y función de activación "softmax", que es la más adecuada para tareas de clasificación multiclase.

- **Tipo de red 3:** consta de cuatro capas densas. La primera capa es una capa densa con 256 unidades y que utiliza como función de activación "ReLU", que incluye un *bias initializer* de 0,1 y *bias regularizer* L2 con un factor de 0,01, seguida de una capa de *BatchNormalization* y una capa de *Dropout* con una tasa del 40 %. La segunda capa es otra capa densa con 128 unidades, también con función de activación "ReLU", el mismo *bias initializer* y *regularizer*, seguida también por una capa de *BatchNormalization* y otra capa de *Dropout* del 40 %. La tercera capa densa tiene 64 unidades con los mismos valores de inicialización y regularización de bias, seguida por una capa de *BatchNormalization* y una capa de *Dropout* del 40 %. La capa de salida para finalizar es una capa densa con el número de unidades según el problema para

el que sea, función de activación "softmax", *bias initializer* de 0,1 y *bias regularizer* L2 con un factor de 0,01, útil para tareas de clasificación multiclase.

- **Tipo de red 4:** consta de cuatro capas densas. La primera capa es una capa densa con 256 unidades y que utiliza la función de activación "ReLU", que aplica un *kernel initializer* "He normal" y un *kernel regularizer* L2 con un factor de 0,01, seguida de una capa de *BatchNormalization* y una capa de *Dropout* con una tasa del 40 %. La segunda capa es otra capa densa con 128 unidades y activación "ReLU", donde también se aplica un *kernel initializer* "He normal" y *kernel regularizer* L2, seguida de una capa de *BatchNormalization* y una capa de *Dropout* del 40 %. La tercera capa densa tiene 64 unidades, también con *kernel initializer* "He normal" y *kernel regularizer* L2, seguida de una capa de *BatchNormalization* y una capa de *Dropout* del 40 %. La capa de salida es una capa densa con el número de unidades según el problema para el que sea, función de activación "softmax", *kernel initializer* "He normal" y *kernel regularizer* L2 con un factor de 0,01.

- **Tipo de red 5:** consta de tres capas densas. La primera capa es una capa densa con 256 unidades y que aplica de función de activación "ReLU", que utiliza un *kernel initializer* "He normal" y un *bias initializer* constante de 0,1, además de *kernel* y *bias regularizer* L2 con un factor de 0,01 para ambos, seguida de una capa de *BatchNormalization* y una capa de *Dropout* con una tasa del 40 %. La segunda capa es otra capa densa con 128 unidades y aplicando funcion de activación "ReLU", que también emplea un *kernel initializer* "He normal", *bias initializer* de 0,1 y *bias regularizer* L2, seguida de una capa de *BatchNormalization* y una capa de *Dropout* del 40 %. La capa de salida es una capa densa con el número de unidades según el problema para el que sea, función de activación "softmax", *kernel initializer* "He normal", *bias initializer* de 0,1 y tanto Bias como *kernel regularizer* L2.

- **Tipo de red 6:** consta de cinco capas densas, cada una seguida de una capa de *BatchNormalization* y una capa de *Dropout* con una tasa del 50 % para evitar el sobreajuste. La primera capa densa tiene 512 unidades y aplica la función de activación "ReLU", la segunda capa densa tiene 256 unidades y tambien aplica activación "ReLU", la tercera capa densa tiene 128 unidades y al igual que las anteriores aplica activación "ReLU", la cuarta capa densa tiene 64 unidades tambien aplicando activación "ReLU", y la quinta capa densa tiene 32 unidades con activación "ReLU". La capa de salida es una capa densa con el número de unidades según el problema para el que sea y función de activación "softmax".

Donde se comenta en la estructura de estas redes sobre que la capa de salida es una capa densa con el número de unidades según el problema para el que sea, esto se refiere a que en este proyecto se consideran tres problemas de clasificación multiclase. El primer problema es evaluar el ganador de los partidos. Las redes neuronales que se entrenen para este problema tendrán la estructura comentada para los anteriores tipos pero sin embargo, en la capa de salida habrá tres unidades, ya que la clase puede tener tres opciones

diferentes. Las opciones son que gana el local, que gana el visitante o que el partido finaliza en empate.

Para los otros problemas, donde se pretende predecir el número de goles de tanto el equipo local como el visitante, las variables a predecir en ambos casos cuentan con diez combinaciones posibles. Estas diez combinaciones son el número de goles del equipo en cuestión, que puede ir desde cero goles a nueve goles. Se ha decidido establecer este rango ya que en los datos obtenidos, todos los goles marcados por los equipos en los partidos obtenidos se encuentran en este rango. Por lo tanto, al entrenar estos tipos de redes neuronales comentadas para estos problemas, en estos casos, la capa de salida debe tener diez unidades.

Con estas combinaciones de parámetros, se pueden obtener 324 combinaciones diferentes que son las que se han evaluado para crear los mejores modelos para predecir tanto el resultado del partido, los goles del local y los goles del visitante. Por lo tanto, con los datos obtenidos, para cada combinación de parámetros, se entrena una red neuronal y se evalúa su exactitud. Finalmente se obtienen los parámetros que se utilizaron en la red neuronal que mejor exactitud ha obtenido. Este proceso se realiza para predecir el ganador del partido, los goles del equipo local y los goles del equipo visitante, obteniendo tres combinaciones de parámetros que se comentan a continuación.

Para optimizar este proceso, se van evaluando las combinaciones diferentes de parámetros por paquetes, pivotando alrededor del optimizador. Es decir, en primer lugar se entrenan los modelos y se calculan sus métricas utilizando todas las combinaciones de parámetros pero manteniendo el optimizador "adam". Después se realiza lo mismo manteniendo el optimizador "SGD" y finalmente manteniendo el optimizador "RMSprop". De esta forma las 324 combinaciones distintas se evalúan en tres paquetes de 108 combinaciones distintas cuyo tiempo de ejecución es bastante menor.

Por otro lado, en las primeras pruebas se ha visto que la calidad de los modelos al variar el número de épocas entre 20, 50 y 80 apenas varia, pero sin embargo, el tiempo utilizado por los modelos con 80 épocas es considerablemente mayor en los otros modelos. Por lo tanto, se ha decidido acotar las opciones de este parámetro en valores más pequeños que tengan menores tiempos de ejecución pero donde el rendimiento no se vea afectado. Los valores finalmente evaluados para este parámetro han sido 20, 30 y 40.

En la Tabla 5.12 se ven las diez redes neuronales con diferentes combinaciones de parámetros con mayor valor de exactitud para el problema de predecir el ganador del partido.

Por lo tanto, respecto al ganador del partido, finalmente se ha seleccionado como el mejor modelo con mayor exactitud, una red neuronal con Tensorflow, en la que se han seleccionado de entre las 324 combinaciones posibles de parámetros, una estructura con 3 capas que corresponde con el tipo de red 5 previamente definido, con descenso de gradiente estocástico como optimizador ("SGD"), con entropía cruzada categórica como función de pérdida, entrenada en 20 épocas con un tamaño de lote de 48 que no incorpora *callbacks*. Esta red ha obtenido una exactitud sobre el conjunto de prueba del 0,560.

Ranking	Tipo de red	Optimizador	Épocas	Tamaño de lote	*Callbacks*	Exactitud
1	5	"SGD"	20	48	No	0,560
2	1	"Adam"	50	80	No	0,558
3	2	"Adam"	80	16	Si	0,557
4	5	"Adam"	20	16	No	0,556
5	1	"SGD"	50	16	No	0,554
6	1	"Adam"	80	16	No	0,553
7	2	"Adam"	80	16	No	0,552
8	2	"Adam"	20	80	No	0,551
9	1	"SGD"	50	16	Si	0,551
10	2	"Adam"	20	48	No	0,549

Tabla 5.12: Valor de la exactitud para las diez estructuras de redes neuronales que mayores valores han proporcionado en esa métrica para predecir el ganador del partido.

Ranking	Tipo de red	Optimizador	Épocas	Tamaño de lote	*Callbacks*	Exactitud
1	6	"SGD"	20	80	No	0,370
2	2	"SGD"	80	48	Si	0,368
3	3	"SGD"	80	48	No	0,366
4	3	"SGD"	80	48	No	0,364
5	6	"SGD"	50	48	Si	0,363
6	3	"SGD"	50	80	No	0,362
7	2	"SGD"	50	16	No	0,361
8	6	"SGD"	20	48	Si	0,361
9	3	"SGD"	80	16	Si	0,359
10	3	"SGD"	50	80	Si	0,359

Tabla 5.13: Valor de la exactitud para las diez estructuras de redes neuronales que mayores valores han proporcionado en esa métrica para predecir los goles del local en el partido.

En la Tabla 5.13 se ven las diez redes neuronales con diferentes combinaciones de parámetros con mayor valor de exactitud para el problema de predecir los goles del local en el partido.

Respecto a los goles del local, finalmente se ha seleccionado como el mejor modelo con mayor precisión, una red neuronal con Tensorflow, en la que se han seleccionado de entre las 108 combinaciones posibles de parámetros, una estructura con 5 capas que corresponde con el tipo de red 6, con descenso de gradiente estocástico como optimizador ("SGD"), con entropía cruzada categórica como función de pérdida, entrenada en 20 épocas con un tamaño de lote de 80 que no incorpora *callbacks*. Esta red ha obtenido una exactitud sobre el conjunto de prueba del 0,370.

Ranking	Tipo de red	Optimizador	Épocas	Tamaño de lote	*Callbacks*	Exactitud
1	4	"Adam"	20	80	No	0,383
2	2	"Adam"	50	16	No	0,376
3	6	"Adam"	20	48	Si	0,374
4	2	"SGD"	50	80	No	0,374
5	1	"SGD"	80	16	No	0,374
6	4	"SGD"	50	16	Si	0,374
7	1	"Adam"	80	48	Si	0,371
8	1	"Adam"	20	48	No	0,370
9	4	"SGD"	80	80	No	0,370
10	2	"SGD"	50	48	No	0,368

Tabla 5.14: Valor de la exactitud para las diez estructuras de redes neuronales que mayores valores han proporcionado en esa métrica para predecir los goles del visitante en el partido.

En la Tabla 5.14 se ven las diez redes neuronales con diferentes combinaciones de parámetros con mayor valor de exactitud para el problema de predecir los goles del visitante en el partido.

Respecto a los goles del visitante, finalmente se ha seleccionado como el mejor modelo con mayor precisión, una red neuronal con Tensorflow, en la que se han seleccionado de entre las 324 combinaciones posibles de parámetros, una estructura con 4 capas que corresponde con el tipo de red 4, con "adam" como optimizador, con entropía cruzada categórica como función de pérdida, entrenada en 20 épocas con un tamaño de lote de 80 que no incorpora *callbacks*. Esta red ha obtenido una exactitud sobre el conjunto de prueba del 0,383.

La exactitud de los modelos sobre los goles de los equipos pueden parecer bajas pero se debe tener en cuenta, que predecir el número de goles exacto de un equipo es más complejo, ya que este valor puede tomar muchos valores diferentes.

6: Resultados y conclusiones obtenidas sobre los datos

6.1. Introducción

En este capítulo se detallan los resultados y conclusiones que se han podido obtener de los datos y de los modelos creados para así poder determinar cuáles son las mejores estrategias de rotación de jugadores y así poder definir conclusiones claras que puedan ayudar a los entrenadores y directivos a tomar decisiones.

6.2. Conclusiones extraídas del análisis previo de los datos

Previo al entrenamiento de los modelos, se ha realizado un análisis de los datos obtenidos para extraer posibles patrones sobre ellos y obtener conclusiones que puedan ayudar a los entrenadores a la hora de la toma de decisiones. Para realizar esto, se ha extraído el valor de cada indicador analizado de forma general para cada equipo en el último partido de la temporada que jugase y después se ha realizado la clasificación ordenando de mayor a menor por este indicador entre todos los equipos de esa liga. Después se ha evaluado en cuantas posiciones difiere la posición de este equipo en esta clasificación por el indicador respecto a la clasificación original ordenando por puntos.

Con estos datos, se ha realizado un mapa de calor de manera que las diferencias negativas se marcasen en rojo y las diferencias positivas se marcasen en verde. Mediante este análisis se pretenden desmentir o confirmar pensamientos que están extendidos de forma generalizada en el mundo de fútbol y que pueden ayudar a extraer conclusiones de los datos, como por ejemplo, los equipos que hacen más cambios introduciendo delanteros, anotan más goles. A continuación, se detalla un ejemplo de cómo se explica y analiza este mapa de calor.

En la Figura 6.11 se muestran los datos para los equipos de LaLiga en la temporada 2024 para la proporción de cambios en la alineación de delanteros en sus partidos en general.

nombre	proporción cambios alineación delantero en general
Real-Madrid	-19
Barcelona	-4
Girona-Fc	-15
Atletico-Madrid	-1
Athletic-Bilbao	-12
Real-Sociedad	5
Betis	4
Villarreal	-5
Valencia-Cf	-6
Getafe	-6
Alaves	0
Sevilla	-2
Osasuna	9
Ud-Palmas	5
Rayo-Vallecano	8
Celta	4
Mallorca	-2
Cadiz	16
Granada	11
Almeria	10

Figura 6.11: Explicación sobre el mapa de calor para las diferencias de posiciones de clasificación.

Aquí se puede observar que el Real Madrid tiene un valor de -19 en verde. Esto se explica porque hay 19 posiciones de diferencia entre su posición por la clasificación por puntos y su posición en la clasificación por este indicador que es la proporción de cambios en la alineación de delanteros en sus partidos en general. Por lo tanto, es un equipo con alto rendimiento ya que ocupa posiciones altas en la clasificación por puntos, pero sin embargo no suele realizar muchos cambios de delanteros en sus alineaciones iniciales. Por el contrario, el Cádiz que tiene un valor de 16 en rojo, se debe a que ocupa una baja posición en la clasificación por puntos, pero por otro lado, ocupa una alta posición en la clasificación por este indicador. Por lo tanto, es un equipo que tiene pocos puntos y un bajo rendimiento y que tiene una elevada proporción de cambios de delanteros en su alineación inicial.

En conclusión, en el análisis de este indicador se aprecia que los equipos de la zona alta de la clasificación por puntos tienen valores bajos en la proporción de cambios de delanteros en la alineación inicial y los equipos de la zona baja de la clasificación por puntos tienen valores altos en la proporción de cambios de delanteros en la alineación inicial.

A continuación, se detallan el resto de las principales conclusiones extraídas que se han podido apreciar de manera generalizada en todas las temporadas de las ligas evaluadas mediante este análisis:

- Los equipos que realizan más cambios en las alineaciones iniciales entre un partido y otro tienden a estar en las posiciones más bajas de la clasificación y por el contrario, los equipos que realizan menos cambios en las alineaciones iniciales entre un partido y otro, tienden a estar en las posiciones más altas de la clasificación. Este efecto se aprecia sobre todo si los cambios en las alineaciones iniciales afectan a los defensas o delanteros.

- Los equipos que realizan más cambios antes del descanso tienden a estar en posiciones más bajas en la clasificación. Esto se puede apreciar sobre todo en la Premier League y Bundesliga. Por otro lado, para todas las ligas, los equipos que hacen menos cambios entre los minutos 61 a 75, tienden a ocupar posiciones más altas en la clasificación.

- Los equipos que hacen más cambios sacando defensas e introduciendo jugadores más ofensivos, suelen ocupar posiciones más bajas en la clasificación. Por otra parte, los equipos que hacen menos cambios sacando delanteros e introduciendo jugadores más defensivos, tienden a ocupar posiciones más altas en la clasificación.

- Respecto a la media de los minutos en la que los equipos realizan los cambios, los equipos de la zona alta de la clasificación tienen valores más bajos en este valor, por lo tanto, de media suelen realizar los cambios antes.

- Los equipos de la zona alta de la clasificación suelen hacer menos cambios de jugadores que han sido amonestados con amarilla.

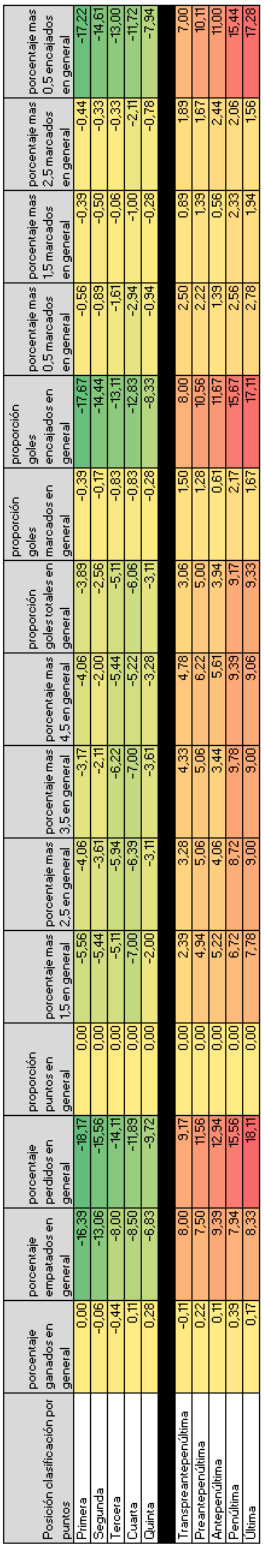
- Finalmente, los equipos de la zona alta de la clasificación suelen tener menores valores en la proporción de cambios que realizan por partido, es decir, en sus partidos no suelen gastar los 5 cambios de los que disponen.

El mapa de calor global agrupando todos los datos de los equipos en las ligas y temporadas consideradas para todos los indicadores analizados se puede ver a continuación en las Figuras 6.12, 6.13 y 6.14. Sin embargo, es importante detallar los siguientes aspectos sobre estos mapas de calor.

- Solo se muestran los datos de las primeras cinco posiciones y de las últimas cinco. En el caso de las últimas cinco se utiliza esa nomenclatura de ultima, penúltima... ya que la Bundesliga tiene menos equipos y por tanto el último de su clasificación ocupa la posición 18 y no la 20 como sucede en LaLiga y la Premier League. Por lo tanto, la última posición, para todas las ligas, se corresponde con el equipo que ocupa la última posición, ya sea la 20 en ligas de 20 equipos (LaLiga y Premier League) o la 18 en la Bundesliga.

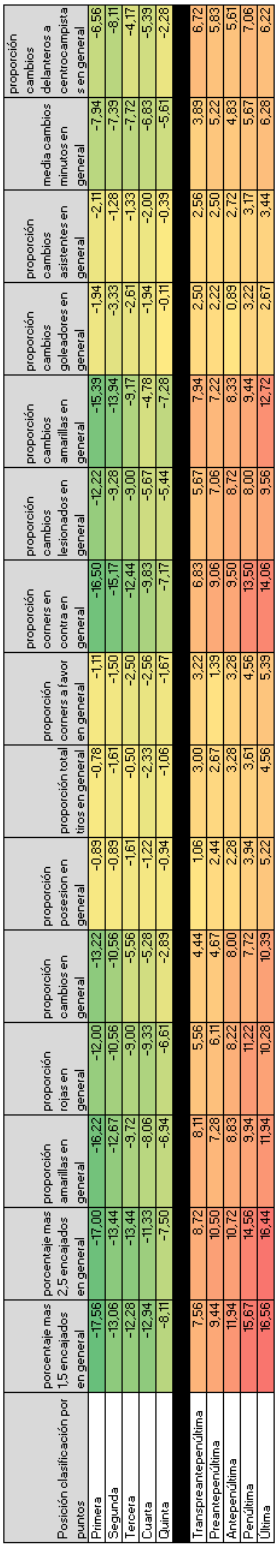
- Cada valor de este mapa de calor global se calcula mediante el promedio de cada valor en esa posición de los datos de cada liga y temporada analizada. Por lo tanto, los datos que se ven en las Figuras 6.12, 6.13 y 6.14, se obtienen al hallar el promedio

de las diferencias de posiciones para cada una de las diez posiciones que se muestran con los datos de las diferentes ligas y temporadas evaluadas. Por ejemplo, para la primera posición del indicador porcentaje de cambios de defensas a centrocampistas, el valor se ha obtenido calculando el promedio con los valores asociados a la diferencia de clasificación del equipo que acabó en primera posición en la clasificación por puntos en cada una de las temporadas y ligas evaluadas para ese indicador.



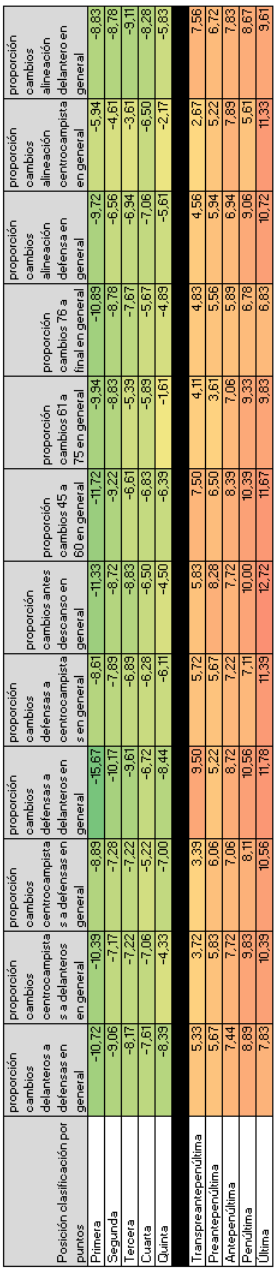
Posición clasificación por puntos	porcentaje ganados en general	porcentaje empatados en general	porcentaje perdidos en general	proporción puntos en general	porcentaje mas 1,5 en general	porcentaje mas 2,5 en general	porcentaje mas 3,5 en general	porcentaje mas 4,5 en general	proporción goles totales en general	proporción goles marcados en general	proporción goles encajados en general	porcentaje mas 0,5 marcados en general	porcentaje mas 1,5 marcados en general	porcentaje mas 2,5 marcados en general	porcentaje mas 0,5 encajados en general
Primera	0,00	-16,39	-18,17	0,00	-5,56	-4,06	-3,17	-4,06	-3,89	-0,39	-17,67	-0,56	-0,39	-0,44	-17,22
Segunda	-0,06	-13,06	-15,56	0,00	-5,44	-3,61	-2,11	-2,00	-2,56	-0,17	-14,44	-0,89	-0,50	-0,33	-14,61
Tercera	-0,44	-8,00	-14,11	0,00	-5,11	-5,94	-6,22	-5,44	-5,11	-0,83	-13,11	-1,61	-0,06	-0,33	-13,00
Cuarta	0,11	-8,50	-11,89	0,00	-7,00	-6,39	-7,00	-5,22	-6,06	-0,83	-12,83	-2,94	-1,00	-2,11	-11,72
Quinta	0,28	-6,83	-9,72	0,00	-2,00	-3,11	-3,61	-3,28	-3,11	-0,28	-8,33	-0,94	-0,28	-0,78	-7,94
Transpreantepenúltima	-0,11	8,00	9,17	0,00	2,39	3,28	4,33	4,78	3,06	1,50	8,00	2,50	0,89	1,89	7,00
Preantepenúltima	0,22	7,50	11,56	0,00	4,94	5,06	5,06	6,22	5,00	1,28	10,56	2,22	1,39	1,67	10,11
Antepenúltima	0,11	9,39	12,94	0,00	5,22	4,06	3,44	5,61	3,94	0,61	11,67	1,39	0,56	2,44	11,00
Penúltima	0,39	7,94	15,56	0,00	6,72	8,72	9,78	9,39	9,17	2,17	15,67	2,56	2,33	2,06	15,44
Última	0,17	8,33	18,11	0,00	7,78	9,00	9,00	9,06	9,33	1,67	17,11	2,78	1,94	1,56	17,28

Figura 6.12: Primera parte del mapa de calor global con las diferencias de posiciones de clasificación.



Posición clasificación por puntos	porcentaje mas 1,5 encajados en general	porcentaje mas 2,5 encajados en general	proporción amarillas en general	proporción rojas en general	proporción cambios en general	proporción posesion en general	proporción total tiros en general	proporción corners a favor en general	proporción corners en contra en general	proporción cambios lesionados en general	proporción cambios amarillas en general	proporción cambios goleadores en general	proporción cambios asistentes en general	media cambios minutos en general	proporción cambios delanteros a centrocampistas en general
Primera	-17,56	-17,00	-16,22	-12,00	-13,22	-0,83	-0,78	-1,11	-16,50	-12,22	-15,39	-1,94	-2,11	-7,34	-6,56
Segunda	-13,06	-13,44	-12,67	-10,56	-10,56	-0,83	-1,61	-1,50	-15,17	-9,28	-13,94	-3,33	-1,28	-7,39	-8,11
Tercera	-12,28	-13,44	-9,72	-9,00	-5,56	-1,61	-0,50	-2,50	-12,44	-9,00	-9,17	-2,61	-1,33	-7,72	-4,17
Cuarta	-12,94	-11,33	-8,06	-9,33	-5,28	-1,22	-2,33	-2,56	-9,83	-5,67	-4,78	-1,94	-2,00	-6,83	-5,39
Quinta	-8,11	-7,50	-6,94	-6,61	-2,89	-0,94	-1,06	-1,67	-7,17	-5,44	-7,28	-0,11	-0,39	-5,61	-2,28
Transpreantepenúltima	7,56	8,72	8,11	5,56	4,44	1,06	3,00	3,22	6,83	5,67	7,94	2,50	2,56	3,89	6,72
Preantepenúltima	9,44	10,50	7,28	6,11	4,67	2,44	2,67	1,33	9,06	7,06	7,22	2,22	2,50	5,22	5,83
Antepenúltima	11,94	10,72	8,83	8,22	8,00	2,28	3,28	3,28	9,50	8,72	8,33	0,89	2,72	4,83	5,61
Penúltima	15,67	14,56	9,94	11,22	7,72	3,94	3,61	4,56	13,50	8,00	9,44	3,22	3,17	5,67	7,06
Última	16,56	16,44	11,94	10,28	10,39	5,22	4,56	5,39	14,06	9,56	12,72	2,67	3,44	6,28	6,22

Figura 6.13: Segunda parte del mapa de calor global con las diferencias de posiciones de clasificación.



Posición clasificación por puntos	proporción cambios delanteros a defensas en general	proporción cambios centrocampistas a delanteros en general	proporción cambios centrocampistas a defensas en general	proporción cambios defensas a delanteros en general	proporción cambios defensas a centrocampistas en general	proporción cambios antes descanso en general	proporción cambios 45 a 60 en general	proporción cambios 61 a 75 en general	proporción cambios 76 a final en general	proporción cambios alineación defensa en general	proporción cambios alineación centrocampista en general	proporción cambios alineación delantero en general
Primera	-10,72	-10,39	-8,89	-15,67	-8,61	-11,33	-11,72	-9,94	-10,89	-9,72	-5,94	-8,83
Segunda	-9,06	-7,17	-7,28	-10,17	-7,89	-8,72	-9,22	-8,83	-8,78	-6,56	-4,61	-8,78
Tercera	-8,17	-7,22	-7,22	-9,61	-6,89	-8,83	-6,61	-5,39	-7,67	-6,94	-3,61	-9,11
Cuarta	-7,61	-7,06	-5,22	-6,72	-6,28	-6,50	-6,83	-5,89	-5,67	-7,06	-6,50	-8,28
Quinta	-8,39	-4,33	-7,00	-8,44	-6,11	-4,50	-6,39	-1,61	-4,89	-5,61	-2,17	-5,83
Transpreantepenúltima	5,33	3,72	3,39	9,50	5,72	5,83	7,50	4,11	4,83	4,56	2,67	7,56
Preantepenúltima	5,67	5,83	6,06	5,22	5,67	8,28	6,50	3,61	5,56	5,94	5,22	6,72
Antepenúltima	7,44	7,72	7,06	8,72	7,22	7,72	8,39	7,06	5,89	6,94	7,89	7,83
Penúltima	8,89	9,83	8,11	10,56	7,11	10,00	10,39	9,33	6,78	9,06	5,61	8,67
Última	7,83	10,39	10,56	11,78	11,39	12,72	11,67	9,83	6,83	10,72	11,33	9,61

Figura 6.14: Tercera parte del mapa de calor global con las diferencias de posiciones de clasificación.

6.3. Conclusiones extraídas al entrenar los modelos

Se ha podido observar que en las predicciones realizadas por el modelo para predecir el ganador del partido, las variables que más repercusión tienen para aumentar la probabilidad de victoria del equipo local son la proporción de este equipo de realizar cambios entre los minutos 61 a 75, la proporción de cambios de defensas a centrocampistas y la proporción de

cambios de centrocampistas a defensas y para el visitante la proporción de este equipo de cambios de centrocampistas a defensas, la proporción de cambios de defensas a delanteros y la proporción de cambios entre el minuto 76 al final. Los equipos con valores más elevados en estos indicadores tienen según el modelo más probabilidad de ganar el partido.

Por otro lado, para el modelo que predice la probabilidad que tiene el equipo local de marcar un determinado número de goles, las variables que tienen más repercusión para aumentar las probabilidades de que el equipo marque más goles son la proporción de este equipo de cambios entre los minutos del 45 al 60, la proporción de cambios de defensas a centrocampistas, la proporción de cambios de defensas a delanteros y la proporción de cambios de centrocampistas a delanteros. Los equipos locales con valores más elevados en estos indicadores tienen según este modelo más probabilidades de anotar un número más elevado de goles.

Para el modelo que predice la probabilidad que tiene el equipo visitante de marcar un determinado número de goles, las variables que más repercusión tienen para aumentar las probabilidades de que el equipo marque más goles son la proporción del equipo de cambios de delanteros en la alineación inicial, la proporción de cambios de centrocampistas en la alineación inicial, la proporción de cambios de defensas a centrocampistas y la proporción de cambios entre los minutos 76 al final. Los equipos visitantes con valores más elevados en estos indicadores tienen según este modelo más probabilidades de anotar un número más elevado de goles.

6.4. Conclusiones y resultados generales

Previo a los modelos, el análisis previo de los datos ha revelado diferentes patrones que pueden ayudar enormemente a los entrenadores a tomar decisiones sobre los jugadores para aumentar el rendimiento del equipo, como evitar realizar muchos cambios en las alineaciones iniciales o evitar realizar muchos cambios antes del descanso, ya que ambos factores en este caso están estrechamente relacionados con los equipos de la zona baja de la clasificación.

Por otro lado, realizar menos cambios en las alineaciones iniciales y menos cambios entre los minutos 61 y 75 se asocia a equipos que ocupan las posiciones más altas de la clasificación, y esto puede ser una buena señal de que estas estrategias ayudan al buen rendimiento del equipo. En cuanto a las posiciones de los jugadores que son afectados por los cambios, realizar cambios ofensivos quitando defensas es una estrategia que se puede apreciar sobre todo en los equipos de la zona baja de la clasificación y por lo tanto no ayuda a su rendimiento. Por otra parte, realizar pocos cambios sacando delanteros e introduciendo jugadores más defensivos es una estrategia utilizada por los equipos de la zona alta de la clasificación y por lo tanto parece que contribuye a su éxito y buen rendimiento.

Sobre los minutos en los que se realizan los cambios, los equipos que de media realizan los cambios antes suelen estar en la zona alta de la clasificación y lo mismo es aplicable a

los cambios que reemplazan jugadores amonestados con amarilla ya que los equipos con menos cambios de este tipo ocupan posiciones más altas.

Finalmente, otro dato bastante significativo y curioso es que los equipos de la zona alta de la clasificación tienen valores más bajos en el número de cambios por partido que hacen, es decir, no gastan todos los cambios de los que disponen. Esto puede contradecir la creencia generalizada de que al realizar más cambios el equipo debería de rendir más porque introduce jugadores totalmente frescos pero parece que, este análisis muestra lo contrario, que conviene mantener los jugadores que estén en el campo y no necesariamente gastar todos los cambios de los que disponen.

7: Conclusiones generales y líneas de trabajo futuras

7.1. Introducción

En este capítulo se desarrollan las conclusiones del proyecto y se revisa si se han conseguido los objetivos que se plantearon al inicio. Finalmente, se proponen diferentes mejoras para realizar en el futuro para seguir desarrollando el proyecto.

7.2. Conclusiones

Las conclusiones de este proyecto destacan la trascendencia y el alcance que provoca la aplicación de diversas técnicas de inteligencia artificial en el contexto del fútbol, específicamente en la predicción del rendimiento de los equipos mediante el análisis de la rotación de jugadores. Este proyecto ha tenido la capacidad de revelar que la implementación de herramientas de inteligencia artificial, como modelos de aprendizaje automático y análisis de datos, pueden ayudar enormemente en la toma de decisiones a los entrenadores y directivos. Se ha podido observar cómo estas tecnologías pueden ofrecer información crucial que tengan una gran repercusión en la toma de decisiones estratégicas de entrenadores y directivos de equipos, permitiéndoles optimizar la rotación de jugadores de manera más precisa y efectiva y por tanto, mejorar su rendimiento.

Además, en este proyecto se destaca la importancia de disponer de conjuntos de datos completos y de calidad para proporcionar a estos modelos de inteligencia artificial de manera adecuada. La recopilación y preparación de datos precisos y relevantes sobre la rotación de jugadores y el rendimiento deportivo se ha establecido como un componente fundamental para el éxito de este proyecto como se ha podido observar.

Este proyecto puede ayudar a destacar la necesidad de desarrollar herramientas y metodologías específicas que faciliten la integración de la inteligencia artificial en la gestión deportiva, lo que implicaría una colaboración conjunta entre expertos en deportes y científicos de datos y que, en muchos casos ya se está realizando.

En última instancia, este proyecto ha pretendido mostrar el potencial de la inteligencia artificial para transformar y mejorar la gestión y el desempeño de los equipos de fútbol analizando los datos sobre la rotación de sus jugadores. Las conclusiones de este proyecto pretenden invitar a continuar investigando y desarrollando este campo, explorando nuevas tecnologías y metodologías que puedan maximizar el impacto positivo de la inteligencia artificial en el mundo del fútbol y así ayudar en todo lo posible a los directivos y entrenadores.

Sin embargo, también se debe tener en cuenta que la naturaleza imprevisible del fútbol es uno de los mayores desafíos para cualquier modelo de inteligencia artificial entrenado para predecir sus resultados. El fútbol es un deporte en el que los resultados pueden variar considerablemente debido a factores aleatorios y circunstancias incontrolables. Eventos inesperados como lesiones de jugadores clave, decisiones arbitrales controvertidas, condiciones climáticas adversas o simplemente un mal día para un equipo o un jugador pueden influir determinantemente en el resultado final de un partido. Estas variables, imposibles de cuantificar y predecir, introducen un grado de aleatoriedad que dificulta la precisión de cualquier modelo predictivo.

Además, el rendimiento de un equipo de fútbol no solo depende de las estadísticas y datos históricos, sino que también esta influenciado por aspectos intangibles que no se pueden calibrar como la moral del equipo, la cohesión entre los jugadores y la estrategia del entrenador. Estos factores humanos y psicológicos son muy difíciles de cuantificar y aún más difíciles de incluir en un modelo de inteligencia artificial de manera correcta. Por ejemplo, un equipo puede tener un rendimiento muy alto en un partido debido a una motivación, como un derbi entre equipos locales o un partido decisivo en un torneo, factores que no se reflejan debidamente en los datos históricos y estadísticas habituales.

Por último, el fútbol es un juego de un número bajo de goles, lo que significa que pequeños errores en la predicción pueden tener un gran impacto en la precisión del modelo. A diferencia de otros deportes con puntuaciones más altas y continuas, en el fútbol, la diferencia entre un gol y ningún gol es capaz de decidir el resultado de un partido. Este margen estrecho de resultados hace que cualquier modelo tenga que ser extremadamente preciso para lograr una alta exactitud. Además, la dinámica táctica de los equipos, que suele variar de un partido a otro en función del rival, introduce otra variabilidad adicional que los modelos predictivos tradicionales habitualmente no son capaces de capturar completamente. En resumen, la combinación de factores aleatorios, humanos y la naturaleza del deporte en sí mismo hace que la predicción de resultados en el fútbol sea especialmente compleja, difícil y sujeta a una baja exactitud en los modelos de inteligencia artificial pero esto no impide que aún así, un correcto análisis de los datos y la aplicación de técnicas de inteligencia artificial, ayuden a tomar mejores decisiones a los entrenadores y directivos que permitan mejorar el rendimiento de los equipos.

7.3. Revisión sobre la consecución de los objetivos

Sobre los objetivos académicos, sí que se han cumplido los objetivos establecidos al iniciar el proyecto.

En primer lugar, en el proyecto se han aplicado los conocimientos sobre *Big Data* adquiridos en el Máster para optimizar las etapas del ciclo de vida de los datos, asegurando así su máxima utilidad y precisión. Inicialmente, se extrajeron los datos en bruto mediante diversos programas de *scraping*. Posteriormente, se llevó a cabo un proceso de limpieza y preparación de datos para que los modelos se pudiesen entrenar con estos datos y para eliminar valores inconsistentes. A continuación, se realizó un análisis de los datos para detectar patrones y conclusiones que pudiesen ser útiles en el proyecto para así poder ayudar a los entrenadores y documentar cuáles son las mejores estrategias de rotación y qué estrategias aumentan el rendimiento de los equipos. Estas técnicas permitieron estandarizar los datos y aumentar su calidad. De esta manera, al aplicar los conocimientos sobre *Big Data* no solo se consiguió una gestión más efectiva de los datos, sino que también proporcionaron conclusiones más útiles para la toma de decisiones estratégicas por parte de los entrenadores sobre como rotar a los jugadores.

En segundo lugar, se han aplicado los conocimientos sobre *deep learning*, redes neuronales y *machine learning* para desarrollar modelos que traten los problemas planteados en este proyecto y proporcionen resultados que ayuden a las personas a tomar mejores decisiones, en este caso a los entrenadores sobre como rotar a los jugadores. El uso de *deep learning*, con redes neuronales profundas, permitió la creación de modelos mediante redes neuronales con diferentes parámetros capaces de aprender y generalizar a partir de grandes cantidades de datos, pudiendo identificar patrones y relaciones. Además al realizar la optimización de los parámetros y de la estructura de estas redes neuronales, se siguió profundizando en los conocimientos sobre ellas lo que conllevó un gran aprendizaje. Además, también se implementaron técnicas de *machine learning* para entrenar diversos algoritmos, utilizando métodos como los árboles de decisión, máquinas de vectores de soporte y bosques aleatorios. En conjunto, el uso de *deep learning*, redes neuronales y *machine learning* permitió la creación de varios modelos muy útiles para la toma de decisiones que pueden ayudar a los entrenadores a aplicar las mejores estrategias para rotar a los jugadores lo que puede marcar una diferencia significativa.

7.4. Líneas de trabajo futuras

En este proyecto se ha analizado el rendimiento de diferentes modelos de inteligencia artificial sobre el desempeño de los equipos de fútbol basándose principalmente en los datos sobre la rotación de sus jugadores. Sin embargo, por la naturaleza del proyecto, debido a que es un proyecto académico, no se ha podido profundizar al máximo en estos aspectos y por tanto a continuación se definen posibles mejoras que puede tener el proyecto en el futuro y que no se han podido realizar en este trabajo.

- **Incorporación de más ligas:** este aspecto podría incrementar la utilidad del sistema desarrollado de manera que sea capaz de ayudar a dirigentes y entrenadores de más clubes y países. Al abarcar más ligas más usuarios podrían utilizar el sistema.
- **Incorporación de más parámetros relacionados con la rotación de los jugadores:** este aspecto podría ayudar a mejorar el rendimiento de los modelos creados y por tanto proporcionar mejores resultados. En este proyecto desarrollado se pretenden utilizar los parámetros y variables más útiles, pero como mejora futura, se podría considerar analizar más parámetros que analicen diferentes datos.
- **Automatizar todo el código para que actualice los datos con los resultados de los últimos partidos:** en este proyecto, de manera inicial, se ha planteado que se deban ejecutar de manera manual los *scripts* para la obtención de los datos de los últimos partidos, pero sin embargo, esta tarea sería importante automatizarla para el futuro.
- **Desarrollar una aplicación web para mostrar los datos obtenidos:** como mejora final, se podría desarrollar una aplicación web que muestre de una forma más amigable los datos obtenidos de los modelos y que puedan ayudar a los entrenadores y directivos.
- **Integración de datos biométricos y físicos:** se podrían incorporar datos biométricos y físicos (frecuencia cardíaca, niveles de fatiga, velocidad, etc.) de los jugadores para afinar las estrategias de rotación basadas en la condición física real de los jugadores. De esta forma, todavía se podrían tomar decisiones más concretas y justificadas.
- **Ayuda para la realización de cambios en directo durante el transcurso del partido:** una importante mejora podría ser que los modelos fuesen capaces de ayudar a los entrenadores a tomar las decisiones sobre los cambios a realizar en el propio partido con avisos y notificaciones para así tomar las mejores decisiones posibles basadas en datos. Sería útil desarrollar simulaciones avanzadas que permitan a los entrenadores evaluar diferentes escenarios de rotación en tiempo real durante un partido, para tomar decisiones más informadas y así mejorar su capacidad en la toma de decisiones.

Apéndices

Apéndice A

Manual de instalación

A.1. Enlace al repositorio

El código desarrollado durante este proyecto se encuentra en el siguiente repositorio:

https://github.com/chemiya/PR-48-TFM

A.2. Despliegue e instalación

Este código se ha desarrollado para que se puedan procesar datos de cualquier temporada y de cualquier liga. Por lo tanto, con la modificación de unos determinados parámetros se pueden obtener, analizar y procesar los datos de cualquier liga o temporada y entrenar los diferentes algoritmos con ellos. Por lo tanto, para el despliegue e instalación del código, en primer lugar, lo más recomendable es descargar el código del repositorio y subirlo a Google Drive para que se pueda ejecutar en Google Colab.

Para descargar el código del proyecto del repositorio que se encuentra en el Apéndice A.1, se puede descargar el .zip con el código del proyecto o clonar el repositorio mediante el comando:

```
$ git clone https://github.com/chemiya/PR-48-TFM
```

Una vez hecho eso, en caso de que se quieran extraer datos de una determinada liga y temporada, se debe de ir en la carpeta *"scraping"* al fichero "Constantes.py" y establecer ahí las ligas o temporadas de las que se quieren extraer los datos.

Una vez hecho eso, se puede ejecutar el fichero "EjecucionGlobal.py". En caso de que se quiera ejecutar este *script* por comandos, el comando es el siguiente:

```
$ python3 EjecucionGlobal.py
```

Después de hacer eso, al cabo de un tiempo, se generarán diferentes csv con los datos extraídos de las páginas web seleccionadas, cada uno recogiendo los datos de una entidad.

Por último, en el caso de que se quieran probar los modelos creados con estos datos, se debe de coger el csv llamado "indicadoresEquipoHistoricoModelo-[liga]-[temporada]" y situarlo en la carpeta "modelos". El entrenamiento de estos modelos se ha realizado en cuadernos de Jupyter Notebook en Google Colab, por lo tanto, como se ha comentado antes, es recomendable subir esta carpeta con el csv y los cuadernos de Jupyter Notebook a Google Drive y ejecutar los cuadernos desde Google Colab.

Bibliografía

[1] Adonaí Vera. Monitoreo del entrenamiento en tiempo real: early stopping y patience. https://platzi.com/clases/2565-redes-neuronales-tensorflow/42851-monitoreo-del-entrenamiento-en-tiempo-real-early-s/. Accessed: 2024-5-31.

[2] Adrián Rodríguez Mira. La inteligencia artificial en el deporte. https://www.tokioschool.com/noticias/inteligencia-artificial-deporte/. Accessed: 2024-5-31.

[3] AWS. ¿Qué es una red neuronal? https://aws.amazon.com/es/what-is/neural-network/. Accessed: 2024-5-31.

[4] Bookdown. Redes neuronales. https://bookdown.org/victor_morales/TecnicasML/redes-neuronales.html. Accessed: 2024-5-31.

[5] Bruce Schoenfeld. El arma secreta del Liverpool: el análisis de datos. https://www.nytimes.com/es/2019/05/29/espanol/liverpool-champions.html. Accessed: 2024-5-31.

[6] datascientest. Beautiful Soup : ¿cómo aprender a hacer web scraping en Python? https://datascientest.com/es/beautiful-soup-aprender-web-scraping. Accessed: 2024-4-23.

[7] datascientest. Cross-Validation : definición e importancia en Machine Learning. https://datascientest.com/es/cross-validation-definicion-e-importancia. Accessed: 2024-5-31.

[8] datascientest. Keras: todo sobre la API de Deep Learning. https://datascientest.com/es/keras-la-api-de-deep-learning. Accessed: 2024-4-23.

[9] datascientest. Pandas : La biblioteca de Python dedicada a la Data Science. https://datascientest.com/es/pandas-python. Accessed: 2024-4-23.

[10] Douglas Vieira. Google Colab: ¿Qué es y cómo usarlo? https://www.hostgator.mx/blog/google-colab/. Accessed: 2024-5-31.

[11] Hugo Cerezo. Neymar ficha por el Paris Saint Germain, que paga los 222 millones de la cláusula. https://www.marca.com/futbol/barcelona/2017/08/03/5981b9d7468aebed368b46a3.html. Accessed: 2024-5-31.

[12] Iberdrola. Descubre los principales beneficios del 'Machine Learning'. https://www.iberdrola.com/innovacion/machine-learning-aprendizaje-automatico. Accessed: 2024-5-31.

[13] IBM. El modelo de redes neuronales. https://www.ibm.com/docs/es/spss-modeler/saas?topic=networks-neural-model. Accessed: 2024-5-31.

[14] Ignacio G.R. Gavilán. Catálogo de componentes de redes neuronales (III): funciones de pérdida. https://ignaciogavilan.com/catalogo-de-componentes-de-redes-neuronales-iii-funciones-de-perdida/. Accessed: 2024-5-31.

[15] Jesús. Optimizadores en deep learning. https://datasmarts.net/es/que-es-un-optimizador-y-para-que-se-usa-en-deep-learning/. Accessed: 2024-5-31.

[16] JOHAN CRUYFF INSTITUTE. Big data, el aliado en la gestión de fichajes y análisis del rendimiento. https://johancruyffinstitute.com/es/blog-es/administracion-del-futbol/big-data-en-la-gestion-de-fichajes-y-analisis-del-rendimiento/. Accessed: 2024-5-31.

[17] Jon Larkin Alonso. ¿Qué es TensorFlow y para qué sirve? https://www.incentro.com/es-ES/blog/que-es-tensorflow. Accessed: 2024-4-23.

[18] José Luis del Olmo Arriaga. El gran negocio del futbol. https://www.theeconomyjournal.com/texto-diario/mostrar/1525487/gran-negocio-futbol. Accessed: 2024-4-23.

[19] Kinsta. ¿Qué Es el Web Scraping? Cómo Extraer Legalmente el Contenido de la Web. https://kinsta.com/es/base-de-conocimiento/que-es-web-scraping/. Accessed: 2024-4-23.

[20] Kristalina Georgieva. La economía mundial transformada por la inteligencia artificial ha de beneficiar a la humanidad. https://www.imf.org/es/Blogs/Articles/2024/01/14/ai-will-transform-the-global-economy-lets-make-sure-it-benefits-humanity. Accessed: 2024-4-23.

[21] LaLiga. LaLiga Beyond Stats. https://www.laliga.com/beyondstats. Accessed: 2024-4-23.

[22] MASTER-DATA-SCIENTIST. SCIKIT-LEARN, HERRAMIENTA BÁSICA PARA EL DATA SCIENCE EN PYTHON. https://www.master-data-scientist.com/scikit-learn-data-science/. Accessed: 2024-4-23.

[23] MEJORCONSALUD. ¿Cómo se usa la inteligencia artificial en el fútbol profesional? https://mejorconsalud.as.com/inteligencia-artificial-futbol-profesional/. Accessed: 2024-4-23.

[24] MIGUEL ÁNGEL DE DIOS. Scrum: qué es y cómo funciona este marco de trabajo. https://www.wearemarketing.com/es/blog/metodologia-scrum-que-es-y-como-funciona.html. Accessed: 2024-5-31.

[25] NAGESH SINGH CHAUHAN. Métricas De Evaluación De Modelos En El Aprendizaje Automático. https://www.datasource.ai/es/data-science-articles/metricas-de-evaluacion-de-modelos-en-el-aprendizaje-automatico. Accessed: 2024-5-31.

[26] NICK HOTZ. Scrum for Data Science. https://www.datascience-pm.com/scrum/. Accessed: 2024-5-31.

[27] OPTA. OPTA DATA. https://www.statsperform.com/opta/. Accessed: 2024-4-23.

[28] ORACLE. Premier League y Match Insights, con tecnología Oracle Cloud: Reimaginando la experiencia de los aficionados. https://www.oracle.com/es/premier-league/. Accessed: 2024-4-23.

[29] PABLO LONDOÑO. Qué es Python, para qué sirve y cómo se usa (+ recursos para aprender). https://blog.hubspot.es/website/que-es-python. Accessed: 2024-4-23.

[30] PLAN DE RECUPERACION DEL GOBIERNO DE ESPAÑA. Qué es la Inteligencia Artificial. https://planderecuperacion.gob.es/noticias/que-es-inteligencia-artificial-ia-prtr. Accessed: 2024-5-31.

[31] RAONA. Los 10 Algoritmos esenciales en Machine Learning. https://raona.com/los-10-algoritmos-esenciales-machine-learning/. Accessed: 2024-5-31.

[32] REDACCIÓN APD. Las ramas de la inteligencia artificial y sus diferentes aplicaciones. https://www.apd.es/tecnicas-de-la-inteligencia-artificial-cuales-son-y-para-que-se-utilizan/. Accessed: 2024-5-31.

[33] RESULTADOSFUTBOL. ResultadosFutbol. https://www.resultados-futbol.com/. Accessed: 2024-4-23.

[34] SAS. Deep Learning Qué es y por qué es importante. https://www.sas.com/es_es/insights/analytics/deep-learning.html. Accessed: 2024-5-31.

[35] Somos Futboleros. Inteligencia artificial predice el campeón de la Premier League. https://onefootball.com/es/noticias/inteligencia-artificial-predice-el-campeon-de-la-premier-league-38801122. Accessed: 2024-4-23.

[36] SPORT. El nuevo Santiago Bernabéu costará tres veces más de lo previsto. https://www.sport.es/es/noticias/real-madrid/nuevo-santiago-bernabeu-costara-tres-99960904. Accessed: 2024-5-31.

[37] Tensorflow. Transferir el aprendizaje y la puesta a punto. https://www.tensorflow.org/guide/keras/transfer_learning?hl=es-419. Accessed: 2024-5-31.

[38] Àlex Calaff. ¿Cuánto dinero ha ganado cada equipo de la Premier League? https://www.sport.es/es/noticias/premier-league/dinero-ganado-equipo-premier-league-88825234. Accessed: 2024-5-31.