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

12 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, 12 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 cuales 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

A **brief** presentation of the topic addressed in the project.

Keywords

keywords separated by commas.

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 [7]. 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 [10].

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 [5].

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, que cambios

realizar y cómo mantener un equilibrio entre la competitividad y la salud de la plantilla. 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 de manera efectiva estas herramientas tendrán una ventaja considerable en la consecución de 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 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 en constante evolución, donde la línea entre el éxito y el fracaso es muy estrecha.

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.

Por último, la motivación detrás de este proyecto también se debe a su 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 e interesante hacia el deporte [8].

- **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 la efectividad en la defensa y el ataque. Estos datos son proporcionados a algoritmos de aprendizaje automático que son capaces de analizar patrones complejos y entrenar modelos predictivos para estimar las probabilidades de resultados de los partidos. De esta manera, la inteligencia artificial proporciona una herramienta poderosa 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 justificadas [12] [15].

- **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 comprender mejor el juego, evaluar el rendimiento de los jugadores y equipos, y tomar decisiones justificadas. Opta se ha convertido en un recurso fundamental en el mundo del deporte para análisis de datos y seguimiento de estadísticas [11].

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 ejecuccion de este proyecto. Estos se dividen en dos categorías distintas, los objetitvos de desarrollo y los objetivos academicos.

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 tecnicas utilizadas para la obtencion de los datos y las tecnologias 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 Conclusiones y lineas de trabajo futuras: en este capítulo se explican las conclusiones finales adquiridas del proyecto junto a las posibles líneas de trabajo futuras a seguir.

2: Objetivos del proyecto

2.1. Introducción

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

2.2. Objetivos de desarrollo

El principal objetivo de desarrollo es crear 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 precisión obtenida.** Sobre todos los modelos evaluados, se deberán seleccionar los que mejor se comporten y se deberá de analizar que precisiones 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.

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 BigData y ponerlos en practica en un proyecto completo. A continuacion 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 seguido en los proyectos de BigData 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: 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 invención de sistemas que son capaces de realizar tareas que generalmente requieren capacidades humanas. Estas tareas abarcan desde el reconocimiento del habla, la toma de decisiones, la traducción de idiomas y el reconocimiento de patrones.

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

La IA ha revolucionado multiples 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. 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 aprender a partir de datos y realizar predicciones o tomar decisiones sin haber sido programadas para llevar esas tareas. A continuación, se detallan los tipos que existen:

1. **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 este área destacan la regresión lineal y logistica, los árboles de decisión y bosques aleatorios y por último, las máquinas de vectores soporte.

2. **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 este área destaca el algoritmo de Kmeans.

3. **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 este área destaca el algoritmo de Q-learning.

En este proyecto principalmente se van a utilizar algoritmos de aprendizaje supervisado ya que el conjunto de datos esta etiquetado. En este proyecto 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 numero de goles.

Los algoritmos de este area asociados a problemas de clasificacion que se utilizan para entrenar los modelos en este proyecto con los datos obtenidos son los siguientes:

1. **Árboles de decisión:** se encargar de dividir iterativamente el conjunto de datos en subconjuntos basados en las características más determinantes, estableciendo una estructura de árbol donde las hojas representan resultados de las decisiones y los nodos representan los atributos.

2. **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 proyectar los datos a un espacio de una dimensión mayor donde las clases sean linealmente separables.

3. **k-Vecinos Más Cercanos (k-NN):** k-NN es un algoritmo supervisado que permite predecir el valor de una nueva muestra en base a los k ejemplos más cercanos en el espacio de características. No contiene una fase de entrenamiento explícita, lo que lo hace simple y eficaz para conjuntos de datos pequeños.

4. **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 del anterior. Los ejemplos más utilizados son XGBoost y LightGBM, que son bastante eficaces y permiten manejar grandes conjuntos de datos con alta dimensionalidad.

5. **Random forest:** es un algoritmo supervisado que genera múltiples árboles de decisión entrenados en diversos subconjuntos del conjunto de datos y/o características. La predicción final la realiza mediante un proceso de agregación donde se utiliza la votación para clasificación, lo que incrementa la precisión y disminuye el sobreajuste. Destaca por ser robusto y eficaz para manejar conjuntos de datos grandes y complejos.

6. **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 en cierto modo 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 se utilizan para evaluar los modelos creados:

1. **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.

2. **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.

3. **Recall:** es la proporción de verdaderos positivos identificados 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.

4. **F1 Score:** es la media armónica entre la precisión y el recall, estableciendo un balance entre ambas métricas. Se calcula como 2 * (Precision * Recall) / (Precision + Recall). Es adecuada cuando se requiere un equilibrio entre precisión y recall.

5. **Matriz de confusion:** 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 cómo el modelo clasifica cada clase, ayudando a realizar el análisis de errores específicos de una clase.

3.4. Deep learning

El aprendizaje profundo 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 analizados. Se ha popularizado en los últimos años debido a su capacidad para superar a otros algoritmos en tareas como el reconocimiento de imágenes, procesamiento del lenguaje natural y otros campos. A continuación se detallan los componentes de la arquitectura de una red neuronal:

1. **Neuronas artificiales:** simulan el funcionamiento de las neuronas biológicas. Cada neurona recibe varias entradas, las procesa mediante la denominada función de activación y produce una salida.
2. **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.
3. **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 continuacion se detallan los principales tipos de redes neuronales que existen:

1. **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. Su tarea es calcular el gradiente del error con respecto a cada peso aplicando la regla de la cadena, propagando el error desde la salida hacia la entrada.
2. **Redes neuronales convolucionales (CNN):** su arquitectura se divide en capas convolucionales que aplican 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 extraídas, manteniendo 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.
3. **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 diseñada para controlar dependencias a largo plazo mediante la incorporación de celdas de memoria y puertas de entrada, olvido y salida.

En este caso, 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 precisión del modelo. El optimizador ayuda a la red a aprender patrones en los datos realizando iteraciones y ajustes incrementales. Los principales algoritmos de optimizacion para redes neuronales en problemas de clasificación son los siguientes:

1. **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 más frecuente y rápida pero que puede ser ruidosa. Sus ventajas son que es simple y eficiente para grandes conjuntos de datos.

2. **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.

3. **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. Es útil para manejar la tasa de aprendizaje en problemas donde los gradientes son cambiantes.

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 a mejorar la precisión del modelo. Las principales funciones de pérdida para problemas de clasificacion que existen son:

1. **Entropía cruzada (Cross-Entropy Loss):** mide la discrepancia entre las distribuciones de probabilidad predicha y la real, penalizando considerablemente las predicciones incorrectas. Esta función de pérdida es ampliamente utilizada para problemas de clasificación binaria y multiclase, permitiendo ajustar las probabilidades predichas a las verdaderas.

2. **Categorical Cross-Entropy:** es una forma específica de entropía cruzada que se utiliza en clasificación multiclase. Se calcula utilizando la probabilidad predicha para la clase verdadera y es habitual utilizarla en tareas de clasificación de imágenes y texto.

3. **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, penalizando las predicciones alejadas de las etiquetas binarias que realmente son verdaderas.

4. **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 muchas clases.

Las épocas 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.

El batch size en una red neuronal se refiere al número de muestras de entrenamiento 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, pero consume más recursos, mientras que un tamaño de lote menor puede realizar actualizaciones más precisas pero más lentas.

Los callbacks en una red neuronal son funciones personalizables que se activan en momentos específicos durante el entrenamiento, 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 ciertas condiciones. Los callbacks se utilizan para monitorear y mejorar el rendimiento del modelo durante el entrenamiento.

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

1. **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.

2. **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.

3. **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.

4. **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. Validacion cruzada y division de los datos

Por otro lado, para el entrenamiento de los modelos se aplicará la técnica 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.

Al entrenar los modelos, los datos se separan en un conjunto de entrenamiento, prueba y validacion. Dividir los datos en estos 3 conjuntos es crucial para evaluar y mejorar el rendimiento de un modelo. El conjunto de entrenamiento con un 70 % de los datos se utiliza para ajustar los parámetros del modelo, mientras que el conjunto de validación con un 15 % de los datos se emplea para ajustar los hiperparámetros y evitar el sobreajuste. Finalmente, el conjunto de prueba con un 15 % 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 capitulo se detalla que tecnica se utiliza para la extraccion de los datos y que tecnologias 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 [14] mediante *scraping*.

El *scraping* [6] 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 [13]. Una de las bibliotecas más utilizadas en Python es BeautifulSoup 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 elementos específicos y que permitan extraer datos de manera automatizada [1]. Por otro lado, 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 [9]. Además, Python ofrece la biblioteca pandas, que 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 [3].

- **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. Ofrece una sintaxis simple y una abstracción de alto nivel que permite crear modelos complejos de manera rápida, 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. 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 [2].
- **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. Esta tecnología destaca por su capacidad para trabajar con inmensos volúmenes de datos y su eficiencia en la ejecución en variedad de plataformas. TensorFlow ofrece una amplia 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 para proyectos de inteligencia artificial y aprendizaje automático en diversas industrias. Esta tecnología se utiliza en este proyecto

para la creación de modelos más avanzados que pueden tener un rendimiento elevado sobre los datos proporcionados [4].

- **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. Colab ayuda a la colaboración en tiempo real, permitiendo compartir y editar notebooks con diferentes usuarios. Además, se integra perfectamente con Google Drive para el almacenamiento y la gestión de archivos.

5: Aspectos relevantes del desarrollo del proyecto

5.1. Introducción

En este capitulo se recogen los aspectos más interesantes del desarrollo del proyecto, donde se documenta desde la metodologia aplicada para el desarrollo del proyecto, describiendo los pasos a seguir y el alcance que se espera que tenga el proyecto. Por otro lado, tambien se detalla la planificacion del proyecto, cual es el modelo de los datos, que operaciones de transformacion y limpieza han sido necesarias y finalmente se describe la implementacion de los archivos 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 teorica y diferentes redes neuronales utilizando los parametros tambien comentados. Los modelos se entrenarán y ajustarán utilizando los datos que hemos obtenido previamente, y se evaluará su rendimiento utilizando kas métricas comentadas.

Esta fase permitirá detectar los modelos más eficaces y precisos para predecir el rendimiento deportivo basado en la rotación 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, se definirá la metodología y 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 rotacion 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.

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 y no 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 eficacia y precisión en la predicción del rendimiento de los equipos. Después de esto, se seleccionará el mejor modelo y se evaluará su rendimiento en la actualidad aplicándolo sobre partidos que estén por jugarse.

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 investigacion y se desarrolla el nucleo del proyecto. Para finalizar, el proyecto continua como Trabajo de Fin de Master, donde se sigue profundizando y expandiendo el trabajo realizado previamente.

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.

La duracion 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 limite de finalizacion es el 28 de junio de 2024.

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

5.5. Modelo de los datos

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

A continuacion, se realiza una breve descripcion de cada entidad:

5.6. Limpieza y transformación de los datos

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

- **Eliminar datos sobre substituciones de jugadores no detectados:** se han eliminado los registros de datosJugadoresPartidos donde no se ha podido extraer la posicion sobre el jugador sustituido. Esto ha sucedido un con apenas 3 jugadores en todas las ligas y temporadas evaluadas y por lo tanto el numero de registros afectados en minimo.
- **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 unicamente consideran los partidos previos de los equipo 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 como se comporta ese equipo.

- **Eliminacion de ids:** para preparar los datos para entrenar los modelos, se han eliminado tanto el id del partido asociado como el id unico del dato para cada registro con los datos de los indicadores para un partido.
- **Transformacion de la clase:** 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 numero de goles del local o el numero de goles del visitante, aplicando one-hot para que las redes neuronales puedan utilizar estos datos.
- **Normalizacion 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 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.

5.7. Implementación

El codigo del proyecto se ha ejecutado en una maquina virtual proporcionada por la Escuela Todo el codigo del proyecto se ha dividido en diferentes carpetas. A continuacion se detalla la finalidad de cada una estas carpetas y sus archivos:

6: Resultados y conclusiones obtenidas sobre los datos

6.1. Introducción

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

6.2. Conclusiones extraidas del analisis previo de los datos

Previo al entrenamiento de los modelos, se ha realizado un análisis de los datos obtenido para extraer posibles patrones sobre ellos. 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 futbol y que pueden ayudar a extraer conclusiones de los datos, como por ejemplo, los equipos que hacen más cambios introduciendo delanteros, anotan más goles. En la siguiente tabla 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:

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 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.

6.3. Conclusiones extraidas al entrenar los modelos

A la hora de entrenar los modelos, para los datos de cada temporada, se han eliminado los datos asociados a los partidos anteriores a la jornada 10, ya que hasta entonces no se ha considerado que existan datos suficientes sobre los partidos previos que han jugado los equipos y por lo tanto los indicadores y porcentajes pueden tener valores extremos. Esto se debe a que para cada partido para el cálculo de los indicadores solo se tienen en cuenta los partidos previos del equipo en esa temporada.

Después de esto, en total, agrupando los datos de las 3 ligas evaluadas en las temporadas comentadas, se han obtenido los datos asociados a 4822 partidos. Con Python y Tensorflow se han entrenado diferentes modelos de Machine Learning e inteligencia artificial. Entre estos modelos se pueden encontrar árboles de decisión, Naive Bayes, bosques aleatorios, regresiones logísticas, SVM y redes neuronales. Los datos utilizados para entrenar los modelos se han dividido en tres conjuntos y se han optimizado los parámetros de cada uno de los modelos.

En las pruebas iniciales se ha apreciado que las redes neuronales con Tensorflow han obtenido una exactitud mayor a los modelos de Machine Learning que se han citado previamente. Por lo tanto, se decidió seguir profundizando sobre estas redes neuronales. Para ello, en la siguiente tabla, 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 de las épocas a analizar a 3 valores diferentes, ya que no existía mucha diferencia entre estos valores. Con estas combinaciones de parámetros, se pueden obtener 108 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.

Respecto al ganador del partido, 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 3 capas, con descenso de gradiente estocástico como optimizador, con entropía cruzada categórica como función de perdida, entrenada en 20 épocas con un tamaño de batch de 32 que no incorpora callbacks. Esta red ha obtenido una exactitud sobre el conjunto de prueba del 56 %.

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 4 capas, con descenso de gradiente estocástico como optimizador, con entropía cruzada categórica como función de perdida, entrenada en 20 épocas con un tamaño de batch de 64 que no incorpora callbacks. Esta red ha obtenido una exactitud sobre el conjunto de prueba del 37 %.

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 108 combinaciones posibles de parámetros, una estructura con 4 capas, con Adam como optimizador, con entropía cruzada categórica como función de perdida, entrenada en 20 épocas con un tamaño de batch de 64 que no incorpora callbacks. Esta red ha obtenido una exactitud sobre el conjunto de prueba del 38 %.

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

6.4. Conclusiones y resultados generales

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.

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

Todo proyecto debe incluir las conclusiones que se derivan de su desarrollo. Éstas pueden ser de diferente índole, dependiendo de la tipología del proyecto, pero normalmente van a estar presentes un conjunto de conclusiones relacionadas con los resultados del proyecto y un conjunto de conclusiones técnicas. Además, resulta muy útil realizar un informe crítico indicando cómo se puede mejorar el proyecto, o cómo se puede continuar trabajando en la línea del proyecto realizado.

7.1. 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 puede ser capaz de revelar que la implementación de herramientas de inteligencia artificial, como modelos de aprendizaje automático y análisis de datos, puede ayudar en la toma de decisiones a los entrenadores y directivos. En este proyecto se va a poder 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.

Además, en este proyecto 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.

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.

En última instancia, este proyecto pretende 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. Los hallazgos 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.

7.2. Líneas de trabajo futuras

En este proyecto se pretenden analizar el rendimiento de diferentes modelos de inteligencia artificial sobre el desempeño de los equipos de futbol basándose 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 pretende 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 pretenden 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 seria 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.

Apéndices

Apéndice A

Plan de Proyecto

Este apéndice presentará el plan de proyecto elaborado para la realización del trabajo. En el caso de trabajos que supongan el desarrollo de software, será sustituído por el Plan de Desarrollo de Software.

A.1. Introducción

A.2. Planificación temporal

A.3. Estudio de viabilidad

Viabilidad económica

Viabilidad legal

Apéndice B

Especificación de Requisitos

Si el TFM comporta el desarrollo de software, en este apéndice se reunirá la Especificación de Requisitos del mismo.

B.1. Introducción

B.2. Objetivos generales

B.3. Catalogo de requisitos

B.4. Especificación de requisitos

Apéndice C

Documento de Diseño

Si el TFM comporta el desarrollo de software, en este apéndice se reunirá el Documento de Diseño asociado al mismo.

C.1. Introducción

C.2. Diseño de datos

C.3. Diseño procedimental

C.4. Diseño arquitectónico

Apéndice D

Documentación del Programador

En este apéndice se incluye la documentación necesaria para que el programador de aplicaciones pueda comprender la estructura de la solución software aportada y para poder modificarla.

D.1. Introducción

D.2. Estructura de directorios

D.3. Manual del programador

D.4. Compilación, instalación y ejecución del proyecto

D.5. Pruebas del sistema

Apéndice E

Documentación de usuario

En este apéndice se incluye la documentación necesaria para que el usuario sepa cómo debe instalar y usar el sistema desarrollado.

E.1. Introducción

E.2. Requisitos de usuarios

E.3. Instalación

E.4. Manual del usuario

Bibliografía

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

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

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

[4] JON LARKIN ALONSO. ¿Qué es TensorFlow y para qué sirve? https://www.incentro.com/es-ES/blog/que-es-tensorflow. Accessed: 2024-4-23.

[5] 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.

[6] 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.

[7] 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.

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

[9] 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.

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

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

[12] 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.

[13] 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.

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

[15] 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.