Desarrollo de un Modelo Predictivo para Estimar la Cantidad de Goles en los Partidos de la Premier League Inglesa

Cifuentes Osorio Ariel[1],
Mogollón Lozano Juan Pablo[2],
Roa Reina Jhon Isidro[3],
Saldarriaga Gutiérrez Andrei Ricardo[4]

Universidad Central
Maestría en Analítica de Datos
Curso de Automatización
Bogotá, Colombia

May 7, 2024

Contents

1 Introducción (Max 250 Palabras) - (*Primera entrega*)

El fútbol, el deporte más popular del mundo, tiene una historia rica que se remonta siglos atrás, con una estimación de al menos 265 millones de seguidores en todo el mundo. Sus orígenes se remontan a China en el siglo III a.C., pero el deporte moderno se formalizó en Inglaterra en el siglo XIX. La creación de la Premier League en Inglaterra marcó un hito en la historia del fútbol inglés, independizándose de la Football League en 1992. Las apuestas deportivas, parte integral de la cultura deportiva, se han practicado durante siglos y se han globalizado en el siglo XXI gracias a Internet. El uso del aprendizaje automático en el análisis y predicción de resultados deportivos, especialmente en el fútbol, ha demostrado ser efectivo. En Colombia, las apuestas deportivas han experimentado un crecimiento significativo en los últimos años. Se pretende desarrollar un modelo predictivo para aprovechar oportunidades en mercados específicos en las apuestas deportivas en Colombia y a nivel mundial.

2 Características del proyecto de investigación que hace uso de Integración y Automatización de Datos para IA (Max 500 Palabras) - (*Primera entrega*)

El fútbol, conocido como el deporte más popular del mundo, tiene una historia rica y diversa que se remonta a siglos atrás. Se estima que al menos 265 millones de personas en todo el mundo siguen este deporte, lo que lo convierte en una pasión global que trasciende fronteras y culturas (Blakemore, 2023).

Los orígenes del fútbol moderno se remontan a China, donde existen registros de deportes similares jugados en el siglo III a.C. Sin embargo, el deporte tal como lo conocemos hoy en día se formalizó en Inglaterra en el siglo XIX, donde se estableció la base para el desarrollo del juego moderno.

Una parte fundamental de la historia del fútbol es la creación de la Premier League en Inglaterra. A finales del siglo XIX, se estableció la liga profesional que, desde 1888 hasta 1992, se denominaba Football League First Division. Sin embargo, a medida que el fútbol ganaba popularidad, los clubes de la English First Division buscaron reestructurar el deporte para adaptarse a las demandas modernas. Así, el 17 de julio de 1991, los clubes firmaron el Acuerdo de Miembros Fundadores para crear la Premier League, independiente de la Football League y The FA. El 20 de febrero de 1992, los 22 clubes renunciaron a la Football League y el 27 de mayo establecieron la Premier League como empresa, marcando un hito en la historia del fútbol inglés (Premier League, s.f.).

Las apuestas deportivas han sido una parte integral de la cultura deportiva durante siglos. Desde la antigua Roma, donde se tienen registros de apuestas en

eventos deportivos, hasta la Inglaterra moderna, donde las apuestas en carreras hípicas eran comunes a finales del siglo XVII, las apuestas han sido una forma de agregar emoción y participación a los eventos deportivos. En el fútbol, las apuestas se han practicado desde los primeros días de la First División en Inglaterra. En 1961, el Gobierno inglés legalizó las casas de apuestas, lo que llevó a un aumento significativo en la industria. Se abrieron 15.000 locales en el Reino Unido, marcando el comienzo de una era de apuestas deportivas legalizadas y reguladas (Smith, 2018).

En el siglo XXI, el negocio de las apuestas se ha globalizado gracias a Internet. Operadores como Betfair, Bwin o Bet365 operan a nivel mundial, brindando a los apostadores acceso a una amplia gama de mercados y eventos deportivos. Hoy en día, hay más de 8000 casas de apuestas en el mundo, la mayoría con sede en paraísos financieros, lo que refleja la naturaleza globalizada de la industria (Hill, 2020).

Con el auge de las herramientas de análisis de datos en el siglo XXI, una de las ramas de la inteligencia artificial que ha alcanzado mayor popularidad en estos tiempos es el aprendizaje automático. Este campo de investigación se ha aplicado de manera creciente en el ámbito deportivo, especialmente en el análisis y la predicción de resultados deportivos (Baraniuk, 2015). Varios estudios han demostrado la eficacia del aprendizaje automático en la predicción de resultados deportivos, especialmente en el fútbol.

Por ejemplo, Douwe Buursma, de la Universidad de Twente en Holanda, realizó un estudio en 2011 enfocado en las apuestas en los partidos de fútbol. Utilizando eventos de partidos pasados y cuatro metodologías diferentes: redes bayesianas, multiclasificación, rotation-forest y Logitboost, logró un 55

En la Universidad de Stanford, se logró un modelo con una confiabilidad del 66

En la Universidad Islámica Azad, se desarrolló un modelo de predicción para el Barcelona con una confiabilidad del 92variables (Universidad Islámica Azad, 2011).

En 2014, Albina Yezus, de la Universidad de San Petersburgo, presentó un modelo utilizando machine learning, en el que tuvo en cuenta variables como la forma, la concentración, la motivación, la diferencia de goles y la historia en enfrentamientos. Logró un porcentaje de confianza entre el 55.8

Un aspecto importante a considerar en las apuestas deportivas es que un evento de fútbol tiene tres resultados posibles: victoria del equipo local (1), empate (x) y victoria del equipo visitante (2). Si todos los equipos fueran idénticos, las probabilidades de cada uno de estos resultados serían iguales (1/3, 1/3, 1/3). Sin embargo, debido a la diversidad de factores que diferencian a cada equipo, como la calidad de los jugadores, los entrenadores, las instalaciones o el presupuesto del club, es necesario tener en cuenta estos factores para mejorar la fiabilidad de los modelos predictivos. Por ejemplo, explorar mercados que

tengan 2 resultados posibles (SI/NO o Cumple/No cumple), como el total de goles anotados por los equipos, podría ofrecer mejores resultados que los obtenidos en estudios anteriores que superan el 60

En el caso colombiano, las apuestas deportivas han experimentado un crecimiento significativo en los últimos años. En 2017, a través del Decreto 167 de 2017, se aprobó la utilización del internet como canal de comercialización de las apuestas permanentes, lo que marcó un hito en la industria en el país. Para el año 2020, se estimaba que había 17 operadores autorizados en Colombia, con un total de 2.505.934 de cuentas de jugadores inscritos en diferentes páginas y entregas de premios por 2.7 billones de pesos. En 2022, se estimaba que existían ocho millones de cuentas activas y un movimiento de más de 26 billones de pesos en el mismo año, siendo el fútbol el principal motor de ingresos para las casas de apuestas en el país (Gómez, Valiente et al., 2020).

Teniendo en cuenta lo anterior, se pretende desarrollar un modelo predictivo utilizando análisis estadísticos y modelos como Random Forest o Modelo Bayesiano para aprovechar oportunidades en mercados específicos, como el número de goles, tiros de esquina, tarjetas, entre otros, de un equipo en particular. Este enfoque busca mejorar la fiabilidad de los modelos predictivos y aprovechar las oportunidades que ofrece el mercado de las apuestas deportivas en Colombia y a nivel mundial.

2.1 Titulo del proyecto de investigación (Max 100 Palabras) - (*Primera entrega*)

Desarrollo de un Modelo Predictivo para Estimar la Cantidad de Goles en los Partidos de la Premier League Inglesa"

2.2 Objetivo general (Max 100 Palabras) - (*Primera entrega*)

Desarrollar un sistema de integración y automatización de datos para aplicaciones de Inteligencia Artificial que posibilite el análisis predictivo de resultados en partidos de la Premier League, con el propósito de aumentar la precisión en las apuestas deportivas

2.2.1 Objetivos especificos (Max 100 Palabras) - (*Primera entrega*)

- Recopilar datos históricos de las últimas cinco temporadas de la Premier League, incluyendo resultados de partidos, estadísticas de equipos y jugadores, y cualquier otra información relevante
- Integrar los datos recopilados en una base de datos y diseñar un esquema de almacenamiento eficiente para facilitar su acceso y consulta.
- Desarrollar procesos de ETL (Extracción, Transformación y Carga) para automatizar la recopilación, limpieza y carga de datos en la base de datos.

- Explorar y evaluar diferentes algoritmos de aprendizaje automático, como Random Forest, Redes Neuronales y Regresión Logística, para construir modelos predictivos
- Validar los modelos mediante pruebas con datos históricos y comparar su desempeño con otros métodos de predicción.

2.3 Alcance (Max 200 Palabras) - (*Primera entrega*)

El objetivo principal de este proyecto es desarrollar un modelo predictivo para la Premier League que abarque varios mercados de apuestas, como el número de goles anotados, tiros de esquina y tiros a puerta, tanto para el equipo local como para el visitante. Este modelo busca utilizar análisis estadísticos avanzados y técnicas de aprendizaje automático, como Random Forest o Modelos Bayesianos, para mejorar la precisión en las predicciones y aprovechar las oportunidades en los mercados de apuestas deportivas. Se espera que este enfoque permita identificar patrones y tendencias en los datos históricos de la Premier League, lo que podría traducirse en ventajas competitivas para los apostadores al brindarles información valiosa para tomar decisiones más fundamentadas en sus apuestas.

2.4 Pregunta de investigación (Max 100 Palabras) - (*Primera entrega*)

¿Cuál es la mejor manera de integrar y procesar los datos históricos de la Premier League de forma eficiente, permitiendo la construcción de modelos de aprendizaje automático precisos y capaces de predecir resultados de partidos con mayor efectividad, en beneficio de las apuestas deportivas?

2.5 Hipotesis (Max 100 Palabras) - (*Primera entrega*)

La integración y automatización de datos de la Premier League para la creación de modelos predictivos busca mejorar la precisión en las apuestas deportivas. Al analizar estadísticas detalladas sobre equipos, jugadores, condiciones de los partidos y tendencias históricas, se evaluará el porcentaje de precisión de acuerdo al comportamiento del modelo. Este enfoque proporcionará a los apostadores una herramienta basada en inteligencia artificial que les permitirá tomar decisiones más informadas. Identificar patrones y tendencias en los partidos ayudará a aumentar sus posibilidades de éxito al apostar en la Premier League, ofreciendo una ventaja competitiva en el mercado de las apuestas deportivas.

3 Reflexiones sobre el origen de datos e información
(Max 400 Palabras) - (*Primera entrega*)

3.1 ¿Cual es el origen de los datos e información ? (Max 100 Palabras) - (*Primera entrega*)

Se empleará una base histórica que abarca las últimas cinco temporadas de la Premier League, la cual incluye variables como la fecha del partido, los equipos locales y visitantes, los goles anotados por cada equipo, entre otras.

Además, se utilizarán técnicas de web scraping mediante Python para actualizar semanalmente los resultados de la liga en curso. Esto permitirá realizar comparaciones entre las predicciones realizadas y los resultados reales, lo que contribuirá a evaluar la precisión del modelo de predicción en tiempo real.

3.2 ¿Cuales son las consideraciones legales o éticas del uso de la información? (Max 100 Palabras) - (*Primera entrega*)

Los datos empleados en este proyecto son accesibles en la web y no están sujetos a restricciones para su consulta, por lo que no se anticipan problemas legales relacionados con su uso.

Es importante destacar que los resultados de las predicciones no están dirigidos a un público menor de 18 años. Además, se promoverá el uso responsable de la información, ya que las apuestas pueden conllevar riesgos de adicción.

3.3 ¿Cuales son los retos de la información y los datos que utilizara en Integración y Automatización de Datos para IA? (Max 100 Palabras) - (*Primera entrega*)

El proyecto enfrenta desafíos significativos en la gestión de la información y los datos. Uno de los retos principales es asegurar la calidad y la integridad de los datos utilizados, ya que la precisión de los modelos predictivos depende en gran medida de la calidad de la información de entrada. Además, la variedad de fuentes de datos disponibles, que incluyen bases de datos en línea, feeds en tiempo real y registros históricos, requiere un esfuerzo considerable para integrar y normalizar los datos de manera coherente. La velocidad de procesamiento también es crucial, ya que se necesita la capacidad de procesar grandes volúmenes de datos y actualizar los modelos en tiempo real para mantener su relevancia. Por último, la seguridad de los datos es un aspecto fundamental, dado que se manejarán datos sensibles y privados, por lo que se deben implementar medidas sólidas de protección y cumplir con las regulaciones de privacidad vigentes. Superar estos desafíos requerirá una combinación de tecnologías avanzadas, como herramientas de inteligencia artificial y análisis de datos, así como la colaboración de expertos en el dominio del fútbol y las apuestas deportivas.

3.4 ¿Que espera de la utilización de Integración y Automatización de Datos para IA para su proyecto? (Max 100 Palabras) - (*Primera entrega*)

Con la asistencia de la Inteligencia Artificial, se busca desarrollar un modelo predictivo automatizado para eventos deportivos de la Premier League. Este modelo permitirá monitorear los resultados en tiempo real, lo que mejorará nuestra capacidad para realizar análisis en tiempo real y tomar decisiones informadas en las apuestas deportivas. Además, la automatización de los procesos nos permitirá ahorrar tiempo y recursos, ya que no será necesario realizar tareas manuales repetitivas. En resumen, esperamos que la integración de la Inteligencia Artificial en nuestro modelo predictivo nos permita mejorar la precisión y eficiencia en las apuestas deportivas en la Premier League.

4 Diseño de integración y Automatización de Datos para IA (Diagrama) (*Primera entrega*)

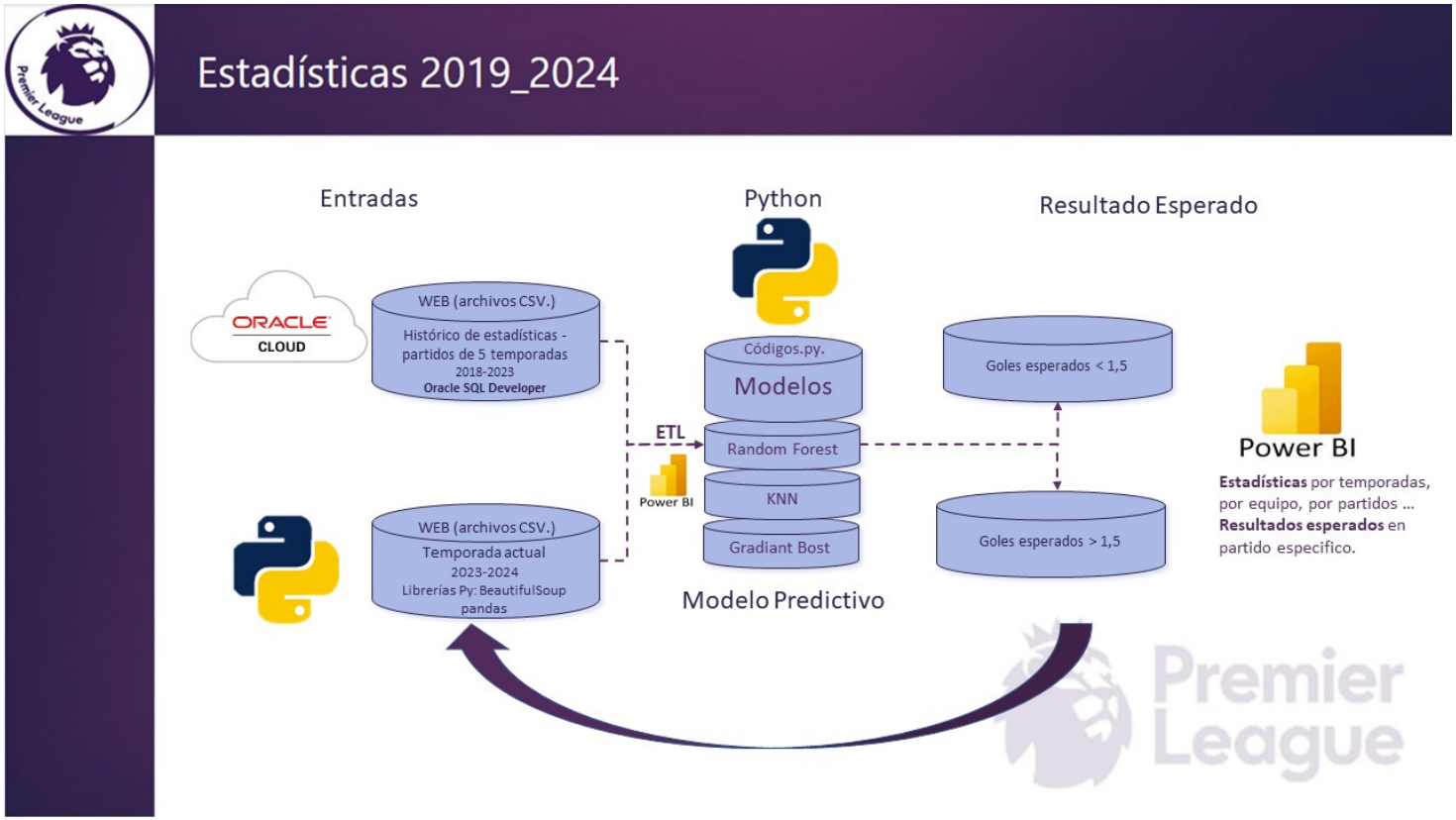

Figure 1: Esquema de trabajo realizado.

5 Integración de Datos (*Segunda entrega*)

La fuente de los datos históricos de las últimas 5 temporadas de la Premier League fue la página https://fbref.com y la extracción de los datos se realizó mediante Python utilizando las siguientes librerías:

1) Request: Esta librería nos permitió hacer solicitudes HTTP de la página web con el fin de interactuar con los recursos que se disponen en ella.

2) BeautifulSoup: Esta librería nos permitió hacer un análisis HTML con el fin de hacer la extracción de la información necesaria de la página web, en este caso, los datos de los partidos de la Premier League desde el año 2019 hasta el año 2023.

3) Pandas: Esta librería nos permitió hacer la recopilación de la información extraída, para convertirla en un Data Frame exportable en formato ".xlsx".

Gestión de Base de Datos

Para la creación de la Base de Datos, seleccionamos Oracle Developer como nuestro Sistema Manejador de Bases de Datos (SMBD), teniendo en cuenta que puede hacer gestión de grandes volumnes de datos, cuenta con una seguridad robusta, posee herramientas de desarrollo integradas, cuenta con soporte para procedimientos almacenados y automatización y tiene compatibilidad con diversas interfaces de programación.

Teniendo en cuenta lo anterior, diseñamos nuestro modelo de datos conforme al siguiente diagrama:

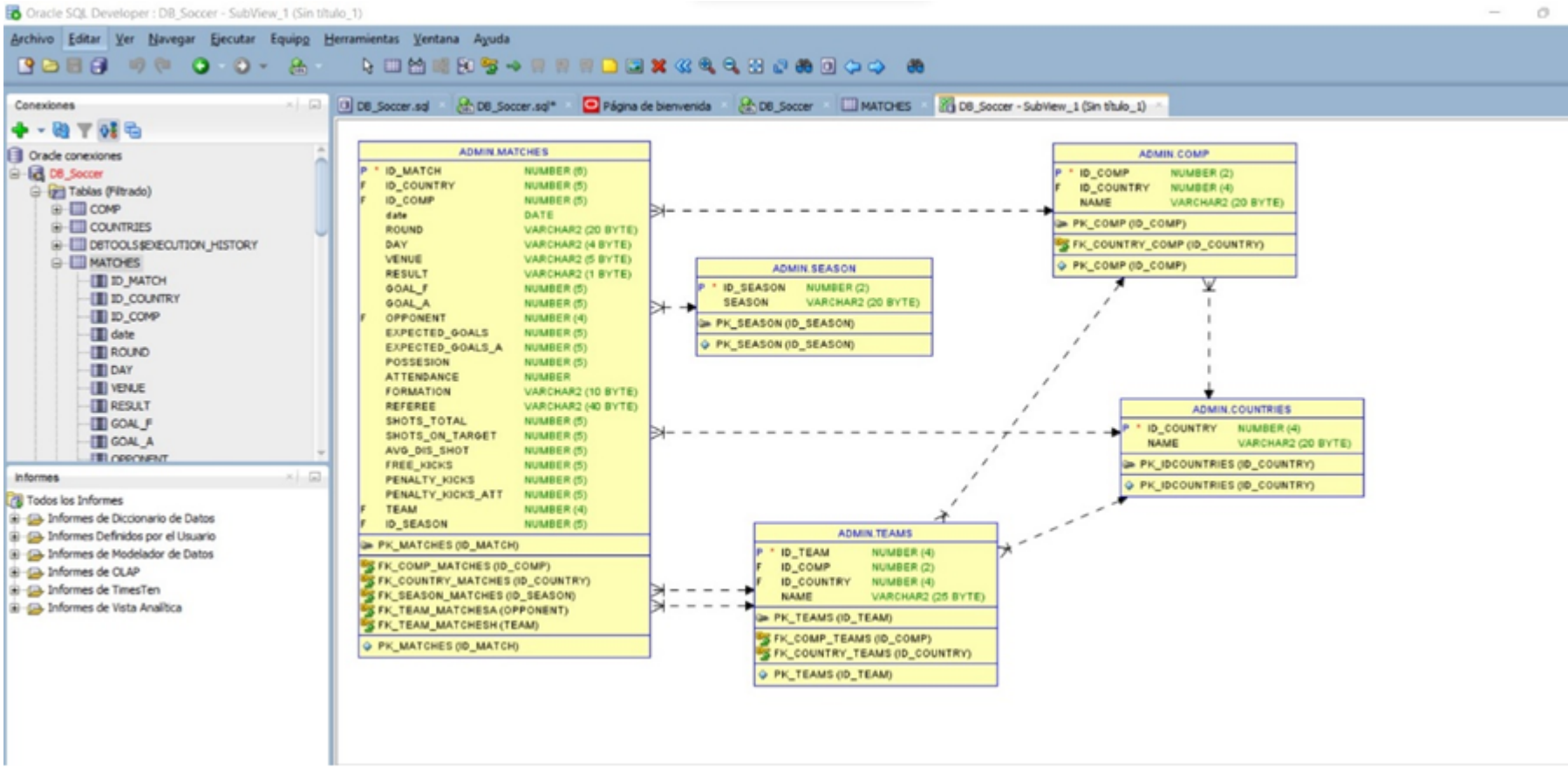

Figure 2: Esquema de la base de datos generada en Oracle.

De nuestro modelo de datos, destacamos las siguientes entidades:

1) Matches: Esta tabla recopila la información de los partidos que se jugaron en la Premier League, desde el mes de agosto de 2018 hasta el mes de junio de 2023. Cuenta con información relevante como lo es; Equipo, oponente, condición de localía, número de goles a favor, número de goles en contra, tiros al arco, tiros libres, posesión de balón, entre otros. 2) Teams: Esta tabla contiene la lista de equipos que participaron en la Premier League, durante las fechas anteriormente, indicadas.

Una vez se cargaron los datos en Oracle Developer, creamos una cuenta en Oracle Cloud, para administrar nuestra base de datos. Teniendo en cuenta que los datos descritos anteriormente, corresponden a datos históricos que no van a sufrir cambios, constituimos la base de datos como Data Warehouse.

Esta base de datos nos servirá para el entrenamiento de nuestro modelo de Machine Learning, del cual se hablará más adelante.

Datos de la Liga Actual

Haciendo uso de las librerías Request, BeautifulSoup y Pandas de Python, también extraeremos la información de los partidos correspondientes a la Premier League que se encuentra en curso. Teniendo en cuenta que cada semana se juegan nuevos partidos de la liga en curso, el Script de Python será el que automatice la extracción de la información de los nuevos partidos, para ser incorporados a nuestra base de datos.

6 Automatización de Datos (*Segunda entrega*)

La automatización de datos es un componente esencial en el desarrollo y la optimización de estrategias en diversos campos, y las apuestas deportivas no son la excepción. En este proyecto de apuestas deportivas, se busca implementar un sistema automatizado que utilice datos históricos de la Premier League para predecir resultados en la temporada actual.

El proceso de automatización comienza con la recopilación de una base histórica que abarca las últimas 5 temporadas de la Premier League. Esta base de datos proporciona un contexto valioso sobre el rendimiento de los equipos en diferentes condiciones y contra distintos oponentes a lo largo del tiempo, base de datos que se compila y almacena por medio de la aplicación Oracle SQL Developer, permitiendo contar con los datos historicos en diferentes tablas como ya se mecionó en el númeral anterior.

Adicionalmente, se cuenta con una base de datos en tiempo real que registra el desempeño de los equipos en la temporada actual. Esta información se

recopila y procesa utilizando Python, un lenguaje de programación ampliamente utilizado en análisis de datos y machine learning.

Posteriormente y empleando la aplicación de Power BI se realiza el proceso de ETL (Extracción, transformación y carga), importando los datos de todas las temporadas, la limpieza y transformación para luego relacionar las tablas de cada temporada y por ultimo consolidar en una sola tabla , la cual se exportará a un archivo CSV para ser empleado en cada uno de los siguientes modelos predictivos: Random Forest, K-Nearest Neighbors (KNN), y Gradient Boosting. Modelos que se implementarán en Python para generar las predicciones sobre los resultados de los partidos en la temporada actual.

El ciclo de automatización se completa con la integración de estos modelos predictivos en un flujo de trabajo continuo. A medida que se actualizan los datos en tiempo real, los modelos se recalibran automáticamente para mejorar su precisión y adaptarse a las tendencias emergentes en la temporada actual de la Premier League.

7 IA (*Segunda entrega*)

Se desarrollo el modelado con el algoritmo Random Forest (Breiman, 2001) que es una técnica de aprendizaje supervisado que genera múltiples árboles de decisión sobre un conjunto de datos de entrenamiento,el desarrollo inicial se realizó con datos de las temporadas comprendida entre los años 2018 a 2023.

Las variables utilizadas para el modelo inicial son Partidos Jugados como Local y visitante, Partidos Ganados como Local y visitante, Partidos Empatados como Local y visitante,Partidos Perdidos de Local y visitante, los Goles a Favor Local y visitante, Goles en Contra como Local y visitante, la Diferencia de Gol como Local y visitante , los Puntos obtenidos como Local y visitante, y la realación entre Puntos y partidos Jugados en condición de Local y visitante.

El algoritmo de Random Forest se destaca por su capacidad para manejar grandes volúmenes de datos y su habilidad para identificar patrones complejos en ellos.

Por otro lado, el algoritmo de K-Nearest Neighbors (KNN) utiliza la similitud entre observaciones para realizar predicciones y puede ser útil para identificar patrones basados en la cercanía entre los datos.

Por último, el algoritmo de Gradient Boosting es conocido por su capacidad para mejorar gradualmente la precisión del modelo combinando múltiples modelos más débiles en un modelo más robusto. Esta técnica puede ser especialmente útil para ajustar las predicciones en función de los errores cometidos por modelos

anteriores, mejorando así la precisión general del sistema de predicción.

Estas técnicas de aprendizaje automático son efectivas para analizar conjuntos de datos complejos y grandes dimensiones, permitiendo obtener predicciones precisas y útiles para la toma de decisiones en las apuestas deportivas.

Inicialmente al evaluar el de desempeño del mejor modelo se obtuvieron los siguientes resultados:

R-cuadrado (local): 0.58 R-cuadrado (visitante): 0.56

8 Proximos pasos (*Tercera entrega*)

9 Lecciones aprendidas (*Tercera entrega*)

10 Bibliografía

Baraniuk, C. (2015). Rise of the AI sports coach. New Scientist, 227(3035). doi: 10.1016/S0262-4079(15)31025-3

Breiman, L. (2001). Random Forests. Machine Learning, 45, 5-32 https://doi.org/10.1023/A:1010933404324

Blakemore, E. (2023, 13 de junio). ¿Dónde surgió el fútbol? Esto dicen los arqueólogos. National Geographic. https://www.nationalgeographic.es/historia/donde-surgio-el-futbol-esto-dic

Canela Ribas, J. (2023, 24 de enero). Estimación de resultados deportivos mediante modelos lineales generalizados. Recuperado de https://diposit.ub.edu/dspace/bitstream/2445/19842

Gómez Figueroa, N., Valiente Ortegón, K. S., Beltrán Suárez, A., Preciado Guerrero, H. M. (2020). Análisis de la influencia de la publicidad en el aumento de la ludopatía por apuestas deportivas. Esp. en Gerencia de Mercadeo. Recuperado de https://repository.universidadean.edu.co/bitstream/handle/10882/10040/BeltranAngelica2020.

Mariño, S. L. (2018). Modelos de predicción de partidos de fútbol en las ligas Española e Inglesa utilizando árboles de clasificación y redes bayesianas. Recuperado de https://repositorio.uniandes.edu.co/server/api/core/bitstreams/5b51e818-9e2d-400e-8803-f19d4

Premier League. (s.f.). Origins. Recuperado de https://www.premierleague.com/history/origins

Soto-Valero, C. (2018). Aplicación de métodos de aprendizaje automático en el análisis y la predicción de resultados deportivos [Application of automated learning methods for analyzing and predicting sports outcomes]. Retos, 34, 377-382. Universidad Central "Marta Abreu" de Las Villas (Cuba). Recuperado de https://www.retos.org

XL Semanal. (2019, 28 de mayo). Apuestas deportivas amañadas en partidos de fútbol de Primera División en España. Recuperado el 2 de mayo de 2024, de https://www.xlsemanal.com/conocer/historia/20190528/apuestas-deportivas-amano-partidos-futbol-primera-d