Escuela Internacional de Postgrado

Trabajo fin de Máster

Máster Universitario en Ingeniería de Software: Cloud, Datos y Gestión de TI

Análisis de datos deportivos con Machine Learning

Realizado por
Johnsiel Antonio Castaños Hernández

Dirigido por
José Antonio Troyano Jiménez

Departamento
Lenguajes y Sistemas Informáticos

Sevilla, Septiembre de 2023

Abstract

One of the greatest challenges that humanity has constantly faced is access to information, and ensuring that its processing serves to continuously improve life and the way in which present events influence the future. Machine learning has undoubtedly been providing useful tools and technologies to become data processing in a day-to-day thing, and where the study of information is the main basis for decision making.

The main objective of this work is the study of machine learning in sports, and the subsequent implementation of a use case on data analysis in soccer to predict if a shot becomes a goal, based on the analysis of past events of the same type.

This project takes previous studies to provide a case of use with strong bases, where a tangible improvement is exhibited with respect to other projects of the same type. The final result demonstrates the viability of machine learning and data processing, not only in the sports field, but in any area of society.

Índice general

Índice de figuras

CAPÍTULO 1

Aspectos Introductorios

1.1– Introducción

La evolución del deporte moderno ha estado influenciada por múltiples factores, pero uno de los más importantes sin duda, ha sido la integración efectiva de avances científicos y tecnológicos en las distintas disciplinas deportivas. Este enfoque ha dejado atrás la improvisación en las investigaciones, abrazando la necesidad de describir, comprender, interpretar y explicar teóricamente la realidad deportiva, e incluso predecirla., todo esto requiere el empleo de métodos y herramientas especializadas de conocimiento.

En el contexto deportivo actual, la toma de decisiones se ha convertido en un pilar fundamental para alcanzar resultados competitivos destacados. La complejidad de estas decisiones exige, en muchas ocasiones, la consulta con expertos o especialistas en estrategia deportiva. En este sentido, se han logrado avances significativos en el desarrollo y aplicación de métodos cuantitativos para el análisis de datos deportivos, con un enfoque particular en técnicas estadísticas.

Este campo de estudio está en constante evolución y desempeña un papel esencial en la comprensión de aspectos técnicos y tácticos en el deporte. El análisis del rendimiento deportivo se ha vuelto cada vez más importante en las últimas décadas y se basa en la observación sistemática de los eventos deportivos, incluyendo el análisis específico por jugada.

Sin embargo, el desafío actual no radica tanto en la obtención de datos deportivos, sino en cómo extraer información valiosa de ellos. La gestión automatizada de información y la inteligencia artificial son áreas que están ligadas a los avances tecnológicos en informática y comunicaciones. En particular, el aprendizaje automático, ha emergido como una disciplina prometedora en el ámbito deportivo.

Una de las ventajas más destacadas del aprendizaje automático en el análisis de datos deportivos es su capacidad para superar las limitaciones de los métodos estadísticos tradicionales. Esto ha sido posible gracias a avances significativos en la adquisición y gestión de datos, impulsados por mejoras tecnológicas.

Los métodos de aprendizaje automático se han aplicado con éxito en el fútbol., por ejemplo, en la predicción de resultados de partidos, análisis del desempeño del equipo o la predicción de lesiones del jugador. Sin embargo, el problema actual radica en la capacidad de caracterizar y seleccionar jugadores basándose en los datos disponibles de rendimiento utilizando la máquina.

Con el rápido aumento del volumen de datos de fútbol en formato digital, el uso de métricas específicas para caracterizar y clasificar a los jugadores según sus habilidades ha atraído la atención de entrenadores y científicos de datos.

El uso de herramientas para analizar el rendimiento de los jugadores de fútbol profesionales basándose en los datos disponibles representa una importante ventaja competitiva. En este escenario, queda clara la necesidad de herramientas que puedan buscar efectivamente información en grandes conjuntos de datos de fútbol. La minería de datos es un campo de la informática que se ocupa con el descubrimiento de patrones interesantes en los datos. Un paso importante en el proceso de minería de datos es la aplicación de métodos de aprendizaje automático, que se relaciona con la obtención y gestión conocimiento a partir de datos.

En este informe se hace una revisión de la bibliografía actual que engloba el Machine Learning y la Inteligencia Artificial, analizando los conceptos clave, y su posterior uso en el análisis de datos deportivos, orientado en este caso hacia el fútbol. En este contexto, el objetivo de este trabajo es la exploración de las técnicas de análisis existentes para la determinación de una estrategia que posibilite el desarrollo de pruebas de conceptos para predecir la probabilidad de que un disparo acabe en gol, además de determinar posibles lesiones de jugadores de futbol. Es decir, el objetivo principal es estudiar de forma preliminar la creación de un modelo para el fin mencionado.

Este informe se enmarca en abordar la importancia del machine learning en el deporte como gran fuente de conocimiento, y como aporte hacia el logro de mejores resultados deportivos, así, podrá alinearse con los estándares y requerimientos solicitados.

Realizar estas pruebas de conceptos supone de la interacción de muchos elementos, tales como, mejores estrategias deportivas, estudio de los mercados deportivos, análisis del desempeño, entre otros., que inicialmente podrían no ser abordados a plenitud, y que deja abierta la posibilidad de futuras líneas de trabajo.

Una vez descritos los detalles de implementación de estas pruebas de conceptos, se abordarán los detalles de implementación en python con anaconda, utilizando las mejores prácticas de programación. Finalmente se hace un breve análisis sobre los resultados obtenidos como producto de este proyecto.

Desde el punto de vista técnico, el análisis de datos requiere de un conjunto de etapas donde se asegura el correcto procesamiento de la información, es por ello que llevar a cabo las pruebas de conceptos del dominio trabajado supone del uso de diferentes tecnologías, metodologías y estrategias de implementación que aseguren la obtención de resultados de buena calidad.

Aparte de tratar el Machine Learning en el deporte, y de realizar pruebas de conceptos asociada a esto, se pretende destacar que esto no se limita a temas de rendimiento deportivo, identificación del talento de los atletas o análisis técnico del juego, también aporta al cambio en el lado comercial de las organizaciones deportivas. Por otro lado, se brinda una experiencia deportiva de mejor calidad a los fanáticos, lo que se logra por medio de soluciones de aprendizaje automático.

Por último, se presentan las consideraciones finales, donde se exponen los puntos más relevantes, así como las futuras líneas de trabajo que se pueden seguir.

1.2– Conceptualización

El fútbol es el deporte más popular a nivel mundial, con millones de aficionados apasionados. Es un deporte muy dinámico, con márgenes finos y una gran cantidad de variables. El aspecto más importante de este deporte es marcar goles, y muchos factores afectan el resultado de un intento de gol. El dinamismo del deporte hizo difícil analizar adecuadamente sus complejidades desde el principio, por lo que los objetivos esperados y su modelado son relativamente nuevos.

A la hora de realizar el análisis de un partido viene el término de goles esperado (xG), los cuales pueden verse influenciados por momentos aleatorios y por la "suerte". Los casi fallos, los tiros desviados, los errores de los porteros y las decisiones arbitrales controvertidas pueden por sí solos dictar el resultado final.

En el fútbol profesional la probabilidad de anotar varía mucho de un tiro a otro. El objetivo de las pruebas de conceptos a realizar es poder cuantificar las posibilidades de gol en relación con el lugar del tiro, además de analizar la posible influencia de otros factores.

Tanto el término disparo como gol se verán durante todo el desarrollo y planteamiento de las pruebas de conceptos a realizar; es por ello preciso dar una definición de los mismos en esta etapa del informe.

Un tiro se define como un intento directo de marcar un gol que resulta en un golpe de un jugador al balón hacia la portería contraria. Se cuentan como goles sólo si el portero no está en condiciones de detener el tiro, o si el

Algunos de los factores comunes a la hora de determinar si un disparo se convertirá en un gol son:

- La localización del disparo.
- La cantidad de espacio que tenía el jugador que realizaba el tiro.
- El tipo de juego previo al tiro.
- La zona de origen de la jugada previa al disparo.
- La posición del portero y el número de jugadores que podían bloquear el disparo.
- El número de pases en el movimiento previo al tiro.

- El tiempo del partido y el estado del mismo (si el equipo que disparó estaba ganando, perdiendo o empatando).

1.3– Motivación

Desarrollar nuevas métricas de fútbol avanzadas es clave para medir el rendimiento del equipo y de los jugadores. En el fútbol se han utilizado métricas básicas en el pasado para intentar explicar qué ha ocurrido en un partido, o al menos en parte de la actuación de un equipo o de un jugador. Alguno de estas métricas, como el porcentaje de tiempo en posesión del balón, el número de tiros o la distancia recorrida por los jugadores siguen siendo utilizados, especialmente por los medios deportivos.

El uso de métricas basadas en machine learning es definitivamente el futuro del fútbol, por un lado, porque sin dudas ofrece buenos resultados a la hora de llevar acciones encaminadas a lograr mejoras en los partidos, y por el otro lado, porque permite que el análisis de los datos sea más automatizado, y por ende requiere menos esfuerzo humano.

El papel del Machine Learning en la mejora de la toma de decisiones y la previsión en los deportes, entre muchas otras ventajas, se está expandiendo rápidamente y ganando más atención tanto en el sector académico como en la industria. Sin embargo, para muchos públicos deportivos, profesionales y formuladores de políticas, que no son particularmente modernos ni expertos en Inteligencia Artificial y Machine Learning, la conexión entre estos y los deportes sigue siendo confusa.

Uno de los ejes centrales que motiva la realización de este informe y la posterior validación mediante prueba de conceptos, es poder analizar la forma en que el aprendizaje automático ha impactado en el deporte, y de forma particular en el fútbol, haciendo que cambiemos la forma en que pensamos sobre las estrategias de partido y el análisis del rendimiento de los jugadores, pero sin dejar de lado los grandes desafíos para lograr análisis completos, entre ellos la posible vulneración de la privacidad, confidencialidad y seguridad de los atletas.

En este informe se proporciona un resumen relevante sobre las áreas en las que el aprendizaje automático se ha aplicado a la industria del deporte y a la investigación deportiva. Finalmente, se presentan algunos escenarios hipotéticos de cómo la IA y el Machine Learning podrían dar forma al futuro de los deportes.

Es por ello que este proyecto pretende desarrollar un modelo que permita predecir si un disparo se puede convertir en un gol, basado en datos que permiten evaluar el análisis del rendimiento del propio equipo, del oponente y del jugador. Por otro lado, y sin salir del alcance que se planteará sobre el proyecto, se pretende incluir un modelo que permita determinar la calidad del jugador que dispara, para tener una predicción más precisa sobre rendimiento del equipo.

1.4– Objetivos

Este trabajo tiene como objetivo principal la exploración de la actualidad en el análisis deportivo, así como el diseño e implementación de pruebas de conceptos que permitan

poner a prueba la utilidad del machine learning en el deporte.

Se pretende demostrar que el procesamiento de datos en el dominio tratado es una base de conocimiento amplia y de mucho interés, porque constituye una gran fuente de información y recursos para todos los actores involucrados durante un partido de fútbol. Por otro lado, el dominio tratado es de alto interés por el gran aporte de datos estadísticos, permitiendo que se creen y perfecciones herramientas para lograr mejores resultados deportivos.

Otros objetivos que se persiguen por medio de este trabajo son:

- Proponer una prueba de concepto para el modelado y representación de los datos relativos a que un disparo se convierta en gol, apegado a la realidad de un partido.
- Comprender e identificar el contexto tecnológico detrás del análisis deportivo.
- Documentar y detallar de forma adecuada cómo el machine learning ha mejorado el fútbol.
- Representar los resultados de las pruebas de conceptos realizadas, utilizando las tecnologías y metodologías más adecuadas.

Para alcanzar los objetivos planteados se realizará un análisis exhaustivo de la actualidad del análisis deportivos, identificando las metodologías y estrategias que mejor representen las pruebas a realizar.

Con respecto a la selección del lenguaje de programación e infraestructura tecnológica, la decisión de elegir una tecnología u otra está basada, primero en los conocimientos previos de programación, y segundo en lo adecuado que pueda ser para la realización de un modelo de predicción que se apegue a un correcto análisis de los datos.

Alcanzar los objetivos planteados sin lugar a dudas se verá reflejado en unos resultados de altos estándares con respecto a lo significativo que es este trabajo, así, se podrá medir si se han alcanzado las metas propuestas por medio de diversas pruebas tanto sobre los datos procesados, como en las pruebas de conceptos realizadas.

CAPÍTULO 2

Inteligencia Artificial (IA), Machine Learning y Análisis de datos

La Inteligencia Articial ha llegado para revolucionar la forma en que todo se relaciona, además de introducir grandes cambios en el procesamiento de datos con miras a mejores resultados presentes y futuros.

2.1– Inteligencia Artificial

La inteligencia artificial se refiere a la capacidad de las máquinas y sistemas informáticos para llevar a cabo tareas que, cuando se realizan por seres humanos, generalmente requieren de inteligencia humana. Estas tareas incluyen el aprendizaje, el razonamiento, la resolución de problemas, el reconocimiento de patrones, la comprensión del lenguaje natural y la toma de decisiones.

Se basa principalmente en la idea de que las máquinas pueden ser programadas para imitar la inteligencia humana y, en algunos casos, incluso superarla en ciertas tareas específicas. La IA se ha convertido en un campo de estudio interdisciplinario que abarca la informática, la matemática, la psicología, la neurociencia y otras disciplinas.

El término inteligencia artificial fue adoptado en 1956, pero se ha vuelto más popular hoy día gracias al incremento en los volúmenes de datos, algoritmos avanzados, y mejoras en el poder de cómputo y el almacenaje [6].

La investigación inicial de la inteligencia artificial exploraba temas como la solución de problemas y métodos simbólicos. Luego se comenzó a entrenar computadoras para que imitaran el razonamiento humano básico. Por ejemplo, la Defense Advanced Research Projects Agency (DARPA, Agencia de Proyectos de Investigación Avanzada de Defensa) realizó proyectos para la planificación urbana y de calles. Mucho tiempo antes que Siri, Alexa o Cortana fueran nombres comunes ya se habían creado otros asistentes virtuales.

Este trabajo inicial abrió el camino para la automatización y el razonamiento formal que se ve hoy en las computadoras, incluyendo sistemas de soporte a decisiones y sistemas de búsqueda inteligentes que pueden ser diseñados para complementar y aumentar las capacidades humanas.

2.1.1. Machine Learning

El Machine Learning, o aprendizaje automático, es una rama de la Inteligencia Artificial, que se destaca como una disciplina que permite a las máquinas absorber conocimiento a partir de datos y mejorar su rendimiento en una variedad de tareas sin necesidad de programación explícita. Su esencia se encuentra en la capacidad de descubrir patrones en los datos y utilizar estos patrones para tomar decisiones informadas o realizar predicciones precisas. Esta tecnología se extiende a numerosos campos, incluyendo la visión por computadora, el procesamiento del lenguaje natural y el reconocimiento de voz, transformando radicalmente la manera en que las máquinas interactúan con el mundo que las rodea y con los seres humanos [7].

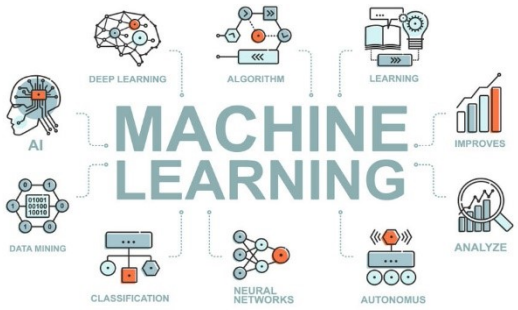


Figura 2.1: Esquematización de Machine Learning.

Las raíces de Machine Learning vienen impulsadas por el intento de modelar matemáticamente las redes neuronales humanas. En 1943, un hito trascendental en esta travesía fue el artículo pionero de Walter Pitts y Warren McCulloch, que se propuso establecer una conexión matemática entre el pensamiento y la toma de decisiones humanas. Este trabajo sentó las bases para la construcción de algoritmos capaces de simular procesos de toma de decisiones inspirados en la inteligencia humana.

El año 1950 marcó otro punto de inflexión en la historia del Machine Learning, cuando Arthur Samuel logró un avance significativo al desarrollar un programa de ordenador capaz de competir en un juego de damas. De manera sorprendente, esta máquina, a pesar de contar con recursos de memoria limitados, demostró habilidades de toma de decisiones utilizando el estratégico algoritmo minimax, sentando así los cimientos del aprendizaje automático [8].

A lo largo de su evolución, el Machine Learning ha madurado para convertirse en una herramienta revolucionaria. Sus algoritmos permiten que el software adquiera conocimiento de forma autónoma, lo que implica la capacidad de adaptación y mejora continua con el tiempo. En la actualidad, el Machine Learning es una disciplina esencial en una amplia gama de campos, desde aplicaciones empresariales hasta investigaciones llevadas a cabo por empresas y universidades. La combinación de algoritmos de aprendizaje automático y modelos de redes neuronales ha llevado el rendimiento de los sistemas informáticos a niveles sin precedentes, impulsando la innovación y la eficiencia en numerosas áreas de aplicación.

Este panorama promete un futuro emocionante donde la Inteligencia Artificial y el Machine Learning seguirán desempeñando un papel central en la transformación de la tecnología y la sociedad. Se vislumbra un mundo donde las máquinas se vuelven aún más inteligentes y capaces de abordar problemas complejos y variados, lo que tendrá un impacto profundo en la forma en que vivimos, trabajamos y nos relacionamos con todo lo que nos rodea. La revolución del Machine Learning continúa, abriendo nuevas fronteras y oportunidades en el vasto y emocionante mundo de la inteligencia artificial.

Tipos de Aprendizaje

En el campo del aprendizaje automático, nos encontramos con dos enfoques fundamentales que son ampliamente reconocidos por su relevancia y utilidad en la resolución de diversas tareas y problemas. Estos dos pilares fundamentales son el aprendizaje supervisado y el aprendizaje no supervisado, cuyo papel es de vital importancia en la construcción de sistemas inteligentes y en el análisis profundo de conjuntos de datos complejos.

1- Aprendizaje supervisado

El aprendizaje supervisado se basa en la construcción de modelos utilizando datos que ya han sido previamente clasificados o etiquetados. Esto significa que el algoritmo tiene acceso a información sobre los datos y se entrena de manera que pueda hacer predicciones coincidentes con los resultados conocidos. El resultado es un modelo capaz de predecir o clasificar nuevos datos que no han sido previamente procesados, basándose en el conocimiento adquirido durante la fase de entrenamiento.

Dentro del aprendizaje supervisado, se distinguen dos técnicas principales:

- **Clasificación:** esta técnica se utiliza cuando se trabajan con datos que tienen respuestas categorizadas, es decir, datos que pertenecen a un conjunto predefinido de categorías.
- **Regresión:** se emplea cuando se trabaja con valores que no se agrupan en categorías, sino que tienen un valor continuo. En este caso, se busca establecer una relación entre las variables explicativas y la variable continua para predecir resultados numéricos.

El proceso de aprendizaje supervisado se divide en dos fases fundamentales, sin importar la naturaleza de los datos: la fase de aprendizaje y la fase de predicción como se logra ver en la imagen.

Proceso de desarrollo del aprendizaje supervisado

En la primera fase, conocida como fase de aprendizaje, se ajusta el algoritmo de manera que se crea un modelo que pueda explicar los resultados en función de los datos disponibles. Se realizan ajustes hasta que el modelo logra hacer predicciones precisas. Una vez que el modelo es válido, se puede usar para predecir resultados con nuevos datos no procesados previamente.

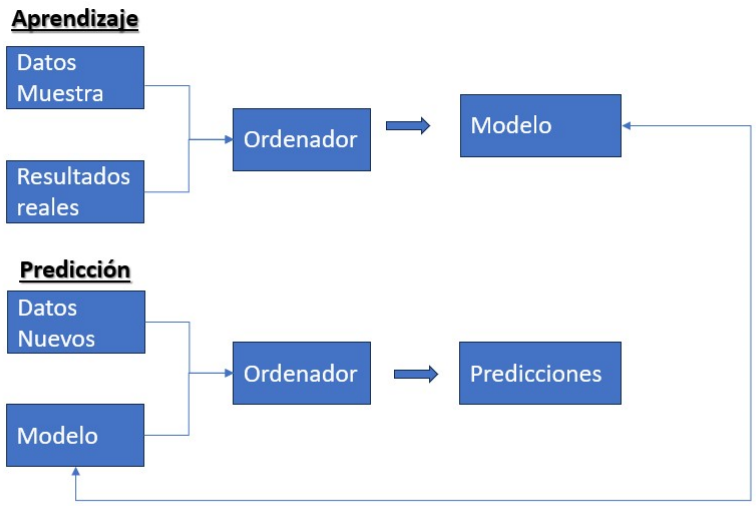


Figura 2.2: Proceso de desarrollo del aprendizaje supervisado.

El objetivo principal del Machine Learning o aprendizaje automático, es crear algoritmos que aprendan y se adapten continuamente. Por lo tanto, los resultados predichos pueden retroalimentarse en el proceso de aprendizaje junto con los datos reales para mejorar y refinar aún más el modelo.

La elección de la técnica y del proceso específico depende de los objetivos de análisis y de la naturaleza de los datos en el aprendizaje supervisado.

2- Aprendizaje no supervisado

En el aprendizaje no supervisado, los datos utilizados no están previamente categorizados ni etiquetados, lo que significa que no se dispone de información adicional aparte de los propios datos en sí. Dado que no se cuenta con conocimiento previo sobre los datos de entrada ni sobre los datos de salida, el éxito del modelo radica en la capacidad de la máquina para identificar patrones, estructuras y relaciones dentro de los datos. El objetivo principal es crear grupos o reglas de asociación que permitan clasificar un nuevo dato en función de la similitud con otros datos [9].

Dentro de las técnicas del aprendizaje no supervisado, se pueden distinguir las siguientes:

- **Clusterización:** consiste en agrupar los datos según sus características sin tener información previa sobre su estructura. Cada grupo, conocido como clúster, está compuesto por objetos que son similares entre sí pero diferentes de los objetos en otros clústeres.

- **Reducción de dimensiones:** los datos suelen estar representados por un gran número de dimensiones, muchas de las cuales pueden estar correlacionadas entre sí. El objetivo aquí es reducir el número de dimensiones explicativas para simplificar el

análisis, conservando las dimensiones principales que capturan la mayor parte de la variabilidad de los datos.

- **Reglas de asociación:** busca identificar patrones que expliquen la co-ocurrencia de características presentes simultáneamente en los datos. Permite establecer con cierto grado de confianza que si una característica se presenta, es probable que otra también esté presente. La elección de la técnica a utilizar depende de los objetivos específicos del análisis y de la naturaleza de los datos.

El aprendizaje no supervisado se centra en descubrir estructuras ocultas y relaciones dentro de los datos sin la necesidad de etiquetas o categorías previas. Las técnicas de clusterización, reducción de dimensiones y reglas de asociación son herramientas poderosas para explorar y comprender la información contenida en conjuntos de datos no etiquetados. Estas técnicas se aplican en una variedad de contextos y campos, desde la segmentación de clientes en marketing hasta la identificación de patrones en datos científicos y empresariales. En la siguiente imagen se logra ver como es el proceso de desarrollo del aprendizaje no supervisado:

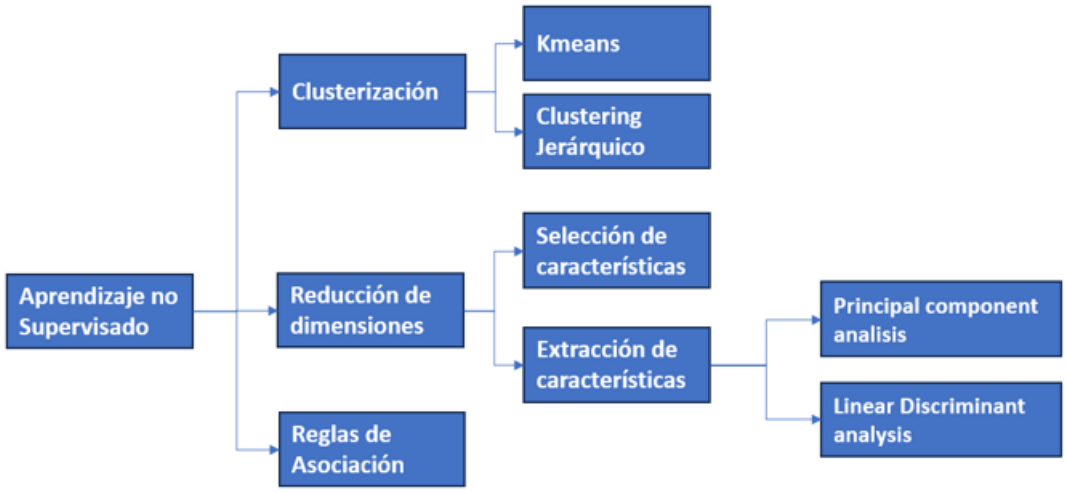


Figura 2.3: Proceso de desarrollo del aprendizaje no supervisado.

3- Aprendizaje por refuerzo

El aprendizaje por refuerzo o reinforcement learning es otro paradigma de machine learning, y se centra en el desarrollo de modelos de toma de decisión que deben ser capaces de maximizar una recompensa a lo largo del tiempo. Generalmente estos sistemas comienzan conociendo poco de su entorno, con el que se familiarizan mediante un proceso exhaustivo de prueba y error [10].

El entrenamiento del modelo equilibra la pura exploración de decisiones inesperadas, potencialmente equivocadas que exponen al modelo a situaciones desconocidas y la optimización de las decisiones en base a las experiencias conocidas. El aprendizaje por refuerzo es particularmente útil en entornos ricos en información de recompensa. Por ejemplo, el

videojuego Pac-Man aumenta la puntuación cada vez que se come un punto, teniendo además muy bien definida la condición de derrota.

En 2013, DeepMind presentó un excelente trabajo donde conseguían superar un gran conjunto de juegos de Atari mediante técnicas de aprendizaje por refuerzo y deep learning. Sin embargo, en entornos tal vez más realistas donde las recompensas están más espaciadas en el tiempo, estos algoritmos tienen más problemas. Por ejemplo, el juego Montezuma's Revenge se convirtió en una bestia negra del aprendizaje por refuerzo, e hicieron falta técnicas inspiradas en la curiosidad4 para poder finalmente dominarlo.

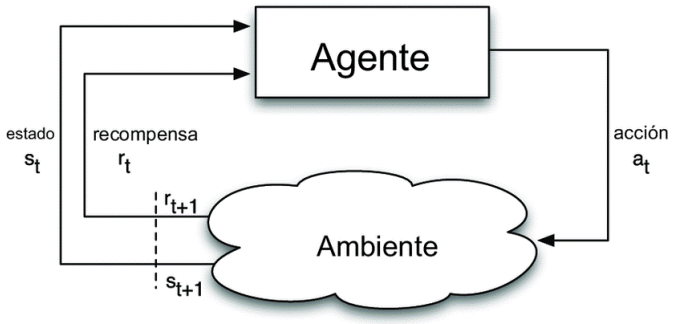


Figura 2.4: Proceso de desarrollo del aprendizaje por refuerzo.

2.1.2. Big Data

El término Big Data engloba conjuntos de datos excepcionalmente grandes o la combinación de múltiples conjuntos de datos que se distinguen por su impresionante tamaño (volumen), su complejidad marcada por una diversidad de fuentes y formatos (variabilidad) y su vertiginoso ritmo de crecimiento (velocidad) [11]. Estas características hacen que gestionar, procesar y analizar estos datos sea un desafío monumental para las tecnologías y herramientas convencionales. Las bases de datos relacionales, las técnicas estadísticas tradicionales y los paquetes de visualización estándar se ven superados por las demandas de Big Data en términos de tiempo y recursos.

A pesar de que el concepto de Big Data como tal es relativamente nuevo, sus raíces se remontan a las décadas de 1960 y 1970, cuando comenzaron a surgir los primeros centros de datos y se desarrollaron las bases de datos relacionales. Desde entonces, el Big Data ha evolucionado y se ha convertido en una fuerza impulsora en la era digital, transformando la manera en que las organizaciones gestionan y aprovechan sus recursos de datos.

Si bien no existe una medida exacta que defina con precisión cuándo un conjunto de datos se convierte en "Big Data", en la actualidad, la mayoría de los expertos y profesionales tienden a considerar como tales a conjuntos de datos que oscilan entre los 30-50 Terabytes hasta varios Petabytes. Estos volúmenes masivos de información se generan constantemente y provienen de diversas fuentes, como registros web, identificación por radiofrecuencia, sensores integrados en dispositivos, maquinaria industrial, vehículos, búsquedas en Internet, interacciones en redes sociales como Facebook, datos de dispositivos móviles como teléfonos inteligentes y tabletas, información de sistemas GPS y registros de centros de

llamadas, entre otros.

La complejidad inherente al Big Data se debe en gran medida a la naturaleza no estructurada de una parte significativa de estos datos. Esta falta de estructura hace que sea más desafiante analizarlos y encontrar patrones significativos. Además, el ritmo acelerado de crecimiento de estos datos agrega una dimensión adicional a la complejidad, ya que las organizaciones deben adaptarse constantemente a la gestión de cantidades cada vez mayores de información.

Para aprovechar efectivamente el potencial del Big Data, en la mayoría de los casos es necesario combinarlo con datos estructurados que generalmente provienen de bases de datos relacionales utilizadas en aplicaciones comerciales convencionales, como sistemas de planificación de recursos empresariales (ERP) o sistemas de gestión de relaciones con el cliente (CRM). Esta fusión de datos estructurados y no estructurados permite obtener una visión más completa y valiosa para la toma de decisiones y el análisis en una amplia variedad de campos, desde el ámbito empresarial hasta la investigación científica.

2.1.3. Datos abiertos (Open Data)

El concepto de Open Data, o Datos Abiertos, constituye una perspectiva esencial en la era digital en la que vivimos. Se trata de la idea fundamental de que ciertos conjuntos de datos deben estar disponibles para cualquier persona sin restricciones de derechos de autor, patentes u otros controles. Esta filosofía de datos abiertos no solo ha ganado relevancia, sino que se ha convertido en una parte integral del mundo digital contemporáneo [12]. Su impacto abarca áreas cruciales como la transparencia gubernamental, la rendición de cuentas, la participación pública y la innovación.

Uno de los aspectos más significativos del enfoque de Open Data es su capacidad para promover la transparencia gubernamental y la rendición de cuentas. Al poner los datos gubernamentales a disposición del público en general de manera abierta y accesible, se permite que los ciudadanos examinen minuciosamente el funcionamiento de sus gobiernos. Este acceso a datos transparentes brinda a los ciudadanos la capacidad de supervisar las acciones y decisiones gubernamentales, y, en última instancia, de exigir responsabilidad por cualquier irregularidad o falta de transparencia.

La tecnología relacionada con los datos abiertos también desempeña un papel fundamental en la participación ciudadana. Al proporcionar información relevante y accesible, se capacita a los ciudadanos para involucrarse de manera más activa en la toma de decisiones que afectan sus comunidades y vidas. Los datos abiertos permiten que las personas formulen preguntas fundamentadas, propongan soluciones basadas en datos y colaboren en la formulación de políticas más informadas y eficaces.

Además de su impacto en la esfera gubernamental, el Open Data también impulsa la innovación. Al abrir conjuntos de datos, se proporciona un recurso valioso para emprendedores, desarrolladores y científicos de datos que pueden utilizar esta información para crear nuevas aplicaciones, servicios y soluciones. Esto no solo estimula la economía digital, sino que también conduce a avances en diversos campos, desde la atención médica y la movilidad urbana hasta la conservación del medio ambiente y la investigación científica.

2.1.4. Datos sin Procesar (Raw Data)

Los datos sin procesar constituyen el punto de partida esencial en la exploración y el análisis de datos, representando la información en su forma más original y sin procesar, tal como se adquiere directamente de diversas fuentes, que pueden incluir sensores, registros, bases de datos, encuestas y redes sociales, entre otras. No obstante, su utilidad inmediata se ve desafiada por una serie de cuestiones cruciales que deben abordarse de manera integral.

En primer lugar, los datos sin procesar a menudo presenta desafíos significativos en términos de calidad y consistencia. Estos pueden manifestarse en forma de errores tipográficos, valores faltantes, duplicados, inconsistencias y ruido en los datos. Como resultado, es imperativo llevar a cabo procesos de limpieza y transformación de datos, que incluyen la detección y corrección de errores, la estandarización de formatos y la imputación de datos faltantes. Este proceso de preparación es crítico para garantizar la calidad y la confiabilidad de los datos en etapas posteriores del análisis.

Además de los desafíos de calidad, la data cruda a menudo carece de contexto y metadatos que describan su origen, significado y estructura. La incorporación de metadatos es esencial para facilitar la interpretación y el uso efectivo de los datos. Estos metadatos actúan como "etiquetas" informativas que ayudan a los analistas y usuarios a comprender mejor los datos y su aplicación potencial.

Otra consideración importante en el manejo de "data cruda" es la privacidad y la seguridad. Dado que los datos pueden contener información sensible o confidencial, es esencial implementar medidas de protección de datos adecuadas para salvaguardar la privacidad de las personas y cumplir con las regulaciones de protección de datos aplicables.

Una vez que los datos crudos se han procesado y preparado adecuadamente, pueden ser utilizados en diversas etapas de análisis, incluido el análisis descriptivo, el modelado predictivo y el análisis exploratorio. La calidad de los datos en estas etapas tiene un impacto directo en la precisión y la validez de los resultados derivados.

La gestión adecuada de "data cruda" también implica la consideración de aspectos como el seguimiento de versiones, especialmente en proyectos de investigación o empresariales a largo plazo, para garantizar la trazabilidad y la reproducibilidad de los resultados en el tiempo [13].

En última instancia, el manejo de "data cruda" implica un enfoque holístico que abarca la preparación de datos, la incorporación de metadatos, la protección de la privacidad, la calidad de los datos y la gestión de versiones. El objetivo final es asegurar que los resultados derivados de estos datos sean precisos, confiables y significativos para respaldar la toma de decisiones informadas y la generación de conocimiento en diversas áreas y aplicaciones.

2.1.5. Almacenes de Datos (Data Warehouse)

Un almacén de datos, o data warehouse, representa un depósito centralizado que alberga y gestiona todos los datos acumulados por los diversos sistemas de una organización.

Este repositorio puede manifestarse tanto físicamente como en forma lógica, y su enfoque principal radica en la recopilación de datos de múltiples fuentes, especialmente con fines analíticos y de acceso facilitado.

Generalmente, un data warehouse encuentra su ubicación en un servidor corporativo o, cada vez con mayor frecuencia, en la nube. Desde allí, se extraen selectivamente datos de diversas aplicaciones de procesamiento de transacciones en línea (OLTP) y otras fuentes para su aprovechamiento por parte de aplicaciones analíticas y para consultas de usuarios.

La finalidad es proporcionar a los ejecutivos y analistas de negocios un entorno donde puedan organizar, comprender y explotar eficazmente sus datos con el objetivo de tomar decisiones estratégicas fundamentadas. La arquitectura de un data warehouse suele dividirse en tres estructuras simplificadas.

La primera es la estructura básica, donde los sistemas operativos y archivos planos suministran datos en su forma cruda y se almacenan junto con metadatos. Los usuarios finales pueden acceder a estos datos para llevar a cabo análisis, generar informes y realizar minería de datos.

Al introducir un área de ensayo entre las fuentes de datos y el almacén en la segunda estructura, se proporciona un espacio para limpiar y preparar los datos antes de que ingresen al almacén principal. Esto permite personalizar la arquitectura del almacén para satisfacer las necesidades de diferentes grupos dentro de la organización.

Finalmente, se puede agregar la tercera estructura, los data marts, que son sistemas diseñados específicamente para atender a una línea de negocio particular. Pueden existir data marts separados para áreas como ventas, inventario y compras, y los usuarios finales tienen la capacidad de acceder a datos de uno o varios data marts de acuerdo con los requisitos de su departamento. Esta arquitectura modular facilita el acceso y el análisis de datos en función de las necesidades de la organización [14].

2.1.6. Segmentación de Datos

La segmentación de datos es una estrategia fundamental en la gestión de la información en el entorno empresarial. Este proceso consiste en la selección y extracción de un conjunto específico de datos de una base de datos de producción, con el propósito de transferirlos a un entorno no productivo. El objetivo principal de la segmentación de datos es reducir el volumen de información, lo que a su vez facilita su manipulación y gestión en entornos de pruebas, desarrollo y análisis.

Para llevar a cabo la segmentación de datos de manera efectiva, se aplican diversos filtros y criterios de selección que permiten acceder a los datos relevantes para un propósito específico. Esto asegura que el conjunto resultante de datos sea coherente y significativo, listo para ser empleado en procesos de prueba de software, análisis estadísticos o cualquier otra aplicación necesaria.

Es importante destacar que la segmentación de datos sigue un enfoque similar al utilizado en la ciencia estadística, donde se utilizan muestras representativas en lugar de

analizar toda la población. Esto garantiza que se trabaje con una porción adecuada de datos que proporcione resultados fiables y representativos sin la necesidad de gestionar grandes volúmenes de información innecesaria.

La implementación de la segmentación de datos conlleva múltiples beneficios para las organizaciones. Uno de los aspectos más destacados es la reducción del tiempo requerido para llevar soluciones de software al mercado, ya que agiliza significativamente los procesos de prueba y desarrollo al trabajar con conjuntos de datos más pequeños y manejables. Además, la disminución en el espacio de almacenamiento necesario en entornos de desarrollo puede ser sustancial, lo que se traduce en un uso más eficiente de los recursos tecnológicos.

Además de los aspectos mencionados, la segmentación de datos también desempeña un papel importante en la gestión de la privacidad y la seguridad de la información. Al seleccionar y manipular sólo los datos necesarios, se reduce el riesgo de exposición de información sensible y se facilita el cumplimiento de las leyes de protección de datos.

2.1.7. Aprendizaje profundo

El aprendizaje profundo, deep learning en inglés, es un campo del aprendizaje automático, machine learning en inglés, que se engloba, a su vez, dentro de la inteligencia artificial.

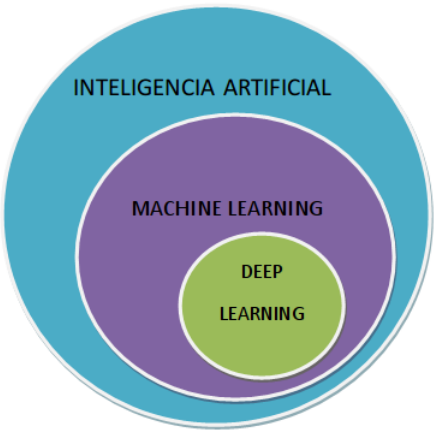


Figura 2.5: Visión del nivel del Aprendizaje Profundo.

El aprendizaje profundo, una subdisciplina del Machine Learning, se basa en algoritmos que operan de manera análoga a las neuronas en el cerebro humano. Este enfoque busca replicar el funcionamiento de las redes neuronales en la construcción de modelos computacionales. En esencia, el aprendizaje profundo instruye a las computadoras para realizar tareas de clasificación directamente a partir de datos como imágenes, texto o sonido, aprendiendo de ejemplos con los que se encuentra.

En el contexto del aprendizaje profundo, los modelos informáticos pueden alcanzar niveles de precisión notables, a veces superando incluso el rendimiento humano en ciertas tareas. Esto se logra mediante la capacitación de estos modelos utilizando conjuntos de

datos amplios y etiquetados, junto con arquitecturas de redes neuronales profundas que cuentan con numerosas capas [15].

El éxito del aprendizaje profundo radica en su habilidad para lidiar efectivamente con grandes conjuntos de datos y en el desarrollo de técnicas de entrenamiento avanzadas. Esto ha llevado a avances significativos en una variedad de aplicaciones y campos, revolucionando la forma en que las máquinas pueden procesar y comprender información.

A diferencia del enfoque tradicional de Machine Learning, donde las características relevantes de los datos deben ser proporcionadas manualmente, el aprendizaje profundo se destaca por su capacidad para extraer automáticamente características clave y aprender las transformaciones necesarias para generar resultados precisos.

Aunque la idea del aprendizaje profundo se planteó por primera vez en la década de 1980, solo en tiempos recientes ha demostrado su utilidad. Esto se debe a dos factores principales: en primer lugar, el aprendizaje profundo requiere enormes conjuntos de datos etiquetados para su entrenamiento, lo que significa que aplicaciones como los automóviles autónomos necesitan millones de imágenes y miles de horas de video. En segundo lugar, el aprendizaje profundo exige una potencia informática significativa, que se ha vuelto más accesible gracias a las GPU de alto rendimiento con arquitectura paralela, así como a la computación en la nube. Estos avances permiten acortar drásticamente los tiempos de entrenamiento de las redes de aprendizaje profundo, pasando de semanas a horas o incluso menos.

2.1.8. Modelos de predicción y Redes Neuronales

Las Redes Neuronales Artificiales (ANN) se han convertido en una piedra angular en el mundo de la inteligencia artificial y el aprendizaje automático. Estas redes ofrecen una aproximación innovadora para abordar problemas complejos en una variedad de campos, desde la visión por computadora hasta el procesamiento del lenguaje natural y la toma de decisiones automatizadas. Para comprender mejor la relevancia y el impacto de las ANN, es esencial explorar en profundidad su funcionamiento y su papel en la resolución de tareas desafiantes.

En el centro de las ANN se encuentra la analogía con el cerebro humano y su sistema nervioso. Aunque las neuronas artificiales no replican la complejidad de las neuronas biológicas, se basan en un principio fundamental: la interconexión. Cada neurona artificial realiza una operación de suma ponderada de sus entradas, seguida de la aplicación de una función de activación que determina su salida. Estas neuronas se organizan en capas o niveles y están interconectadas con un alto grado de conectividad. Las conexiones entre las neuronas están influenciadas por pesos ajustables, lo que permite que la red aprenda de los datos.

Una de las características más impresionantes de las ANN es su capacidad de aprendizaje. A través de algoritmos de aprendizaje, estas redes pueden ajustar sus estructuras y parámetros para minimizar errores en la predicción de datos y mejorar su capacidad para generalizar patrones. Este proceso de aprendizaje es esencial para que las ANN puedan descubrir relaciones ocultas y patrones en conjuntos de datos complejos y diversos.

El aprendizaje profundo, una rama de las ANN, se ha convertido en un enfoque particularmente relevante. Este enfoque implica la creación de redes neuronales profundas con muchas capas intermedias. Estos modelos se entrenan utilizando grandes conjuntos de datos etiquetados y arquitecturas de redes neuronales profundas que pueden comprender la información de manera jerárquica y compleja. El aprendizaje profundo ha demostrado ser especialmente efectivo en tareas de visión por computadora, procesamiento de lenguaje natural y reconocimiento de patrones.

En las últimas décadas, las ANN han ganado prominencia como una tecnología fundamental en la minería de datos. Lo que distingue a estas redes es su capacidad para extraer conocimiento de manera inductiva, basándose en los datos disponibles sin requerir una especificación previa de la estructura funcional y las interacciones. Este enfoque basado en datos ha demostrado ser altamente efectivo para abordar problemas complejos y analizar conjuntos de datos masivos y variados.

2.2– Análisis de datos

El análisis de datos es el proceso de examinar, limpiar, transformar y modelar un conjunto de datos con el objetivo de descubrir información útil, extraer conocimientos y tomar decisiones informadas. Implica la aplicación de técnicas y herramientas estadísticas, matemáticas y de visualización para identificar patrones, tendencias y relaciones en conjuntos de datos. El análisis de datos permite revelar insights, responder preguntas y resolver problemas, ayudando a las organizaciones y personas a comprender mejor el mundo que les rodea, optimizar procesos y tomar acciones basadas en evidencia.

En la actualidad, el uso de las matemáticas y la estadística, junto con la creciente potencia computacional del hardware, principalmente en la nube, ha contribuido a la difusión de metodologías que pretenden replicar la inteligencia humana para extraer información útil en el análisis de datos. Por lo tanto, podemos decir en general que se aprovechan todos los conocimientos lógicos y tecnológicos para apoyar eficazmente las decisiones futuras y comprender mejor lo que ocurrió en el pasado.

2.2.1. Metodologías y tipos de análisis

Las diversas metodologías de análisis de datos se dividen comúnmente en dos grandes categorías: análisis cuantitativos y análisis cualitativos. Los análisis cuantitativos se caracterizan por expresar información en forma numérica, lo que permite realizar cálculos y representar datos en tablas o gráficos. Aunque proporcionan indicaciones sobre cómo categorizar las posibles causas de los problemas o medir su impacto, no ofrecen orientación directa sobre qué problema abordar [1].

- **Análisis Descriptivo:** este tipo de análisis se enfoca en la representación y resumen de los datos disponibles. Se utilizan estadísticas descriptivas, como medias, desviaciones estándar, percentiles y visualizaciones, como gráficos de barras o diagramas de dispersión, para comprender la distribución de los datos y destacar tendencias pasadas. El análisis descriptivo es esencial para proporcionar una base sólida an-

tes de avanzar hacia análisis más avanzados. Por ejemplo, puede revelar patrones estacionales en las ventas o tendencias históricas en el comportamiento del mercado.

- **Análisis Exploratorio:** se emplea cuando se desean descubrir patrones ocultos o identificar posibles áreas de interés en los datos. Aquí, se aplican técnicas avanzadas, como el análisis de clústeres y la reducción de la dimensionalidad, para revelar relaciones no evidentes a simple vista. Este enfoque es especialmente útil en la investigación científica, el descubrimiento de fraudes, la detección de anomalías y la exploración de datos no estructurados, como texto o imágenes.

- **Análisis de Redes:** se concentra en comprender las conexiones entre entidades o nodos en una red, como redes sociales, sistemas de transporte o colaboración empresarial. Este tipo de análisis revela patrones de interacción, influencia y centralidad en la red. Puede utilizarse para identificar líderes de opinión en redes sociales, analizar la estructura de una red logística o comprender las relaciones en una cadena de suministro.

- **Análisis Espacial:** se enfoca en datos con referencia geográfica y busca comprender cómo se distribuyen los fenómenos en el espacio. Esto es esencial en aplicaciones como la planificación urbana, la gestión de recursos naturales y la epidemiología. El análisis espacial puede revelar patrones de concentración geográfica, identificar áreas de alta densidad de eventos y ayudar en la toma de decisiones basadas en la ubicación.

- **Análisis Prescriptivo:** es la etapa más avanzada en el análisis de datos. Aquí, se utilizan modelos predictivos junto con reglas de negocio para recomendar acciones específicas a tomar. Este enfoque es fundamental en la optimización de procesos, la asignación de recursos y la toma de decisiones estratégicas. Por ejemplo, en la logística, puede sugerir rutas óptimas para la entrega de productos, y en el ámbito financiero, puede ayudar a identificar carteras de inversión óptimas.

- **Análisis Predictivo:** se centra en utilizar datos históricos para hacer predicciones sobre eventos futuros. Esto se logra mediante la construcción de modelos de machine learning que pueden identificar patrones y relaciones en los datos. Estos modelos pueden predecir, por ejemplo, la demanda futura de productos, la probabilidad de que un cliente cancele su suscripción o el riesgo de que ocurra un fallo en una máquina. El análisis predictivo es valioso porque permite tomar decisiones proactivas y anticiparse a eventos potenciales.

En este tipo de análisis se basará una de las pruebas de concepto de este trabajo.

2.2.2. Etapas del análisis predictivo

El análisis predictivo, al igual que otros, requiere de un procesamiento exhaustivo de los datos en cuestión, dicho proceso es vital para la obtención de resultados altamente confiables.

Para desarrollar un buen análisis de datos predictivo, de manera efectiva y con la intención de obtener buenos resultados se deben seguir un conjunto de etapas, que se describen en lo adelante.

Esta imagen refleja fielmente el conjunto de etapas que componen el análisis de datos:

Figura 2.6: Etapas del análisis de datos.

Obtención de datos

La obtención de datos para un análisis de datos predictivo en cualquier contexto, es una etapa fundamental en la que se recopilan datos de diversas fuentes para alimentar modelos y sistemas de análisis.

El objetivo principal de esta etapa es recopilar datos de alta calidad y relevantes que sean necesarios para abordar las preguntas o problemas específicos que se desean resolver a través del análisis predictivo. La elección de las fuentes de datos adecuadas y la recopilación precisa son esenciales para garantizar que los modelos predictivos sean precisos y efectivos.

La calidad de los datos y la capacidad para procesarlos adecuadamente son factores críticos en esta etapa. La limpieza, el almacenamiento y la organización de los datos son procesos importantes para garantizar que los datos estén listos para su análisis. Además, se deben considerar aspectos éticos y legales en la obtención y el manejo de datos, especialmente cuando se trata de datos de jugadores y equipos.

En pocas palabras, la obtención de datos es el punto de partida esencial para cualquier análisis de datos predictivo, ya que la calidad y la relevancia de los datos recopilados tienen un impacto directo en la precisión y la utilidad de los modelos predictivos resultantes.

Limpieza y Preparación de datos

La preparación de datos, a menudo pasada por alto pero de suma importancia, desempeña un papel fundamental en el ciclo de análisis de datos. Se trata de un proceso que abarca desde la adquisición de datos en su estado crudo hasta su transformación en un formato adecuado para el análisis. A pesar de que pueda parecer una tarea menos interesante en comparación con la creación de modelos avanzados o la interpretación de resultados, su influencia en la calidad de cualquier análisis no puede subestimarse.

En primer lugar, la validación de datos garantiza que los datos sean precisos y confiables. Esto implica identificar y corregir errores obvios, como valores atípicos, datos faltantes o incoherencias en los registros. La falta de validación de datos podría llevar a conclusiones erróneas o a la toma de decisiones incorrectas basadas en información defectuosa.

El proceso de limpieza de datos, por otro lado, se enfoca en eliminar inconsistencias y ruido de los datos. Esto puede implicar la corrección de errores tipográficos, la estandarización de formatos de fecha o la consolidación de categorías redundantes. La limpieza de datos no solo mejora la calidad de los datos, sino que también facilita la interpretación y el análisis posterior.

La fase de enriquecimiento de datos es otro aspecto esencial. Aquí, se pueden incorporar datos adicionales de fuentes externas para enriquecer la información existente. Por ejemplo, si se están analizando datos demográficos de clientes, agregar datos de ingresos promedio por área geográfica podría proporcionar una perspectiva más completa [2].

Análisis Exploratorio

El Análisis Exploratorio de Datos representa un paso crucial en el proceso de análisis de datos. Va más allá de simplemente examinar los datos y se adentra en el corazón de la exploración, desenterrando información valiosa y revelando patrones ocultos. El an alisis exploratorio de datos utiliza una combinación de técnicas estadísticas y métodos de visualización para sumergirse en los conjuntos de datos, destacando las características más prominentes y detectando detalles que pueden pasar desapercibidos en un vistazo superficial [3].

Una de las principales ventajas del análisis exploratorio de datos radica en su capacidad para ayudar a los científicos de datos a comprender la naturaleza de los datos antes de realizar análisis más avanzados. Al explorar visualmente los datos, se pueden identificar patrones, tendencias y distribuciones de variables clave. Esto permite la detección temprana de posibles problemas, como datos atípicos o valores faltantes, que pueden requerir una atención especial durante el preprocesamiento de datos.

El análisis exploratorio de datos también brinda una visión más profunda de las relaciones entre variables, lo que resulta fundamental para la toma de decisiones informadas. Al comprender cómo interactúan las diferentes variables en un conjunto de datos, los analistas pueden diseñar enfoques de modelado más efectivos y seleccionar las técnicas estadísticas adecuadas. Además, el análisis exploratorio de datos a menudo revela preguntas adicionales y áreas de investigación que pueden no haber sido evidentes inicialmente.

A pesar de que el análisis exploratorio de datos se originó en la década de 1970 con el trabajo pionero de John Tukey, sigue siendo una herramienta esencial en el kit de herramientas de los científicos de datos contemporáneos. En un entorno de análisis de datos cada vez más complejo, esto sigue siendo una práctica estándar para obtener información profunda y bien fundamentada a partir de los datos, lo que es esencial para la toma de decisiones basadas en datos y el avance en la ciencia de datos.

Selección de variables

La selección de características es un paso esencial en la construcción de modelos de machine learning y análisis de datos que tiene como objetivo elegir cuidadosamente un subconjunto de características relevantes de entre todas las disponibles [4]. Esta práctica es fundamental por varias razones que impactan positivamente en la eficacia y la utilidad de los modelos:

- **Simplificación de Modelos:** una de las razones clave para realizar la selección de características es simplificar los modelos. Al reducir la cantidad de variables o características utilizadas en un modelo, este se vuelve más comprensible tanto para los usuarios finales como para los investigadores. La simplicidad es esencial para la interpretación de los resultados, lo que facilita la toma de decisiones basadas en el modelo y permite una comunicación más efectiva de los hallazgos a las partes interesadas.

- **Tiempo de Entrenamiento Más Corto:** en muchos casos, especialmente cuando se trata de conjuntos de datos grandes o complejos, la inclusión de todas las características puede resultar en tiempos de entrenamiento extremadamente largos. La selección de características puede reducir significativamente el número de variables, lo que se traduce en un tiempo de entrenamiento más corto para los modelos. Esto mejora la eficiencia computacional y permite una iteración más rápida en el proceso de modelado.

- **Evitar la Maldición de la Dimensión:** es un fenómeno en el que, a medida que el número de características aumenta, la cantidad de datos requeridos para entrenar de manera efectiva un modelo también debe aumentar exponencialmente. La selección de características aborda este problema al reducir la dimensionalidad del conjunto de datos, lo que puede resultar en modelos más simples y eficientes que aún mantienen un buen rendimiento predictivo.

- **Mejora de la Generalización y Reducción del Overfitting:** la selección de características también puede ayudar a mejorar la generalización de los modelos al reducir el riesgo de sobreajuste (overfitting). Cuando se eliminan características irrelevantes o redundantes, el modelo se vuelve menos propenso a ajustarse demasiado a los detalles específicos del conjunto de entrenamiento y, en cambio, se enfoca en patrones más generales y significativos en los datos, lo que puede mejorar su capacidad para hacer predicciones precisas en nuevos datos no vistos.

Selección del Modelo

La selección de modelos en el análisis predictivo es un proceso crítico que define en gran medida el éxito de un proyecto de data science o machine learning. Implica tomar decisiones estratégicas sobre qué algoritmo o modelo de machine learning utilizar para resolver un problema de predicción específico. Esta elección se basa en varios factores clave que deben ser cuidadosamente evaluados.

En primer lugar, la naturaleza de los datos desempeña un papel fundamental. Los datos pueden ser de diferentes tipos, como datos estructurados, datos de texto o imágenes. Dependiendo de la naturaleza de los datos, ciertos algoritmos pueden ser más adecuados que otros. Por ejemplo, para la clasificación de imágenes, las redes neuronales convolucionales suelen ser la elección preferida, mientras que para la predicción de series temporales, los modelos de series temporales pueden ser más apropiados.

Dependiendo de estos objetivos a lograr, y de lo que se intenta predecir se seleccionará el algoritmo que mejor se adapte a la tarea en cuestión.

El tamaño del conjunto de datos es otro factor a considerar. Algunos algoritmos funcionan mejor con grandes conjuntos de datos, mientras que otros pueden ser más adecuados para conjuntos de datos pequeños. La eficiencia computacional también es importante; algunos algoritmos pueden requerir una gran cantidad de recursos de cómputo, lo que puede no ser factible en todos los casos.

La experiencia y el conocimiento del equipo de trabajo son invaluables en este proceso. Los profesionales con experiencia en machine learning pueden tener una comprensión profunda de los pros y los contras de diferentes algoritmos, lo que facilita la elección adecuada.

Entrenamiento

El proceso de entrenamiento en machine learning es una fase esencial en la creación de modelos inteligentes y predictivos. En esta etapa, se alimenta al algoritmo de machine learning con un conjunto de datos de entrenamiento, del cual aprenderá patrones y relaciones que le permitirán realizar predicciones precisas en el futuro.

Los datos de entrenamiento son cruciales y deben ser representativos del problema que se está abordando. Además, deben incluir la "respuesta correcta" o el resultado deseado, que se conoce como el atributo de destino. Este atributo de destino es lo que el modelo de machine learning intentará predecir a partir de los datos de entrada.

Durante el proceso de entrenamiento, el algoritmo de aprendizaje analiza los datos de entrenamiento y busca patrones en ellos. Estos patrones pueden ser relaciones complejas entre las características de entrada y el atributo de destino. El algoritmo ajusta sus parámetros internos para minimizar la diferencia entre las predicciones que realiza y los valores reales del atributo de destino en el conjunto de entrenamiento.

A medida que el proceso de entrenamiento avanza, el algoritmo de machine learning crea

un modelo de machine learning que actúa como una representación matemática de los patrones aprendidos. Este modelo es lo que se utiliza posteriormente para realizar predicciones en datos nuevos y no vistos. Cuanto mejor sea el proceso de entrenamiento y cuantos más datos de alta calidad se utilicen, más preciso será el modelo resultante en sus predicciones [5].

Evaluación

La evaluación del modelo representa un paso crítico y fundamental en el proceso de análisis predictivo y machine learning. Su objetivo central es medir la capacidad de un modelo de machine learning para realizar predicciones precisas y valiosas en situaciones del mundo real. Esta fase es esencial para garantizar que el modelo sea confiable y eficaz en su aplicación práctica.

La evaluación del modelo se lleva a cabo mediante la utilización de conjuntos de datos de prueba independientes, que no se utilizaron durante el proceso de entrenamiento. Esto simula condiciones del mundo real donde el modelo se enfrentaría a datos desconocidos. La elección de estos conjuntos de datos de prueba es crucial para garantizar que las métricas de evaluación reflejen con precisión el rendimiento del modelo en aplicaciones reales.

Un componente fundamental de la evaluación del modelo es la selección de métricas adecuadas. Las métricas de evaluación varían según el tipo de problema que se esté abordando. Para problemas de clasificación, se utilizan métricas como la precisión, que mide la proporción de predicciones correctas, y la sensibilidad y la especificidad, que evalúan la capacidad del modelo para identificar positivos y negativos. Además, el F1-score y el área bajo la curva ROC proporcionan una comprensión más completa del rendimiento del modelo en situaciones de desequilibrio de clases.

En el caso de problemas de regresión, las métricas como el error cuadrático medio (MSE) y el coeficiente de determinación (R-cuadrado) son comunes. El MSE mide la diferencia entre las predicciones del modelo y los valores reales, mientras que el R-cuadrado proporciona una medida de cuánta variabilidad en los datos es explicada por el modelo.

Estas métricas de evaluación desempeñan un papel crucial en la toma de decisiones relacionadas con el modelo. Permiten comparar diferentes modelos y ajustar parámetros para mejorar su rendimiento. Además, son esenciales para determinar si el modelo cumple con los objetivos de precisión y utilidad en un contexto específico. En última instancia, la evaluación del modelo garantiza que las predicciones generadas sean confiables y valiosas, lo que respalda la toma de decisiones informadas en una variedad de aplicaciones, desde la detección de fraudes hasta la personalización de recomendaciones en línea.

CAPÍTULO 3

Machine Learning en el deporte

Machine Learning (ML) está proporcionado importantes avances en el ámbito deportivo mediante predicciones con alto grado de fiabilidad, evaluaciones objetivas a nivel individual y colectivo, así como mejoras en el rendimiento. El análisis de datos ofrece una visión objetiva de cada jugador tanto en un partido independiente como en toda una temporada.

La adopción de los modelos estadísticos en los deportes se ha vuelto más prominente en los últimos años a medida que las nuevas tecnologías y aplicaciones de investigación están impactando los deportes profesionales en diversos niveles de sofisticación. La amplia aplicabilidad de los algoritmos de aprendizaje automático, combinada con el aumento de la potencia de procesamiento informático, así como el acceso a más y nuevas fuentes de datos en los últimos años, ha hecho que las organizaciones deportivas estén en la búsqueda constante de nuevas aplicaciones y estrategias.

El objetivo primordial sigue siendo hacerlos más competitivos dentro y fuera del campo, tanto en el rendimiento deportivo como en el empresarial. Los beneficios de aprovechar el poder de Machine Learning puede, en ese sentido, tomar diferentes formas, desde optimizar la toma de decisiones técnicas o comerciales hasta mejorar el rendimiento del atleta/equipo, pero también aumentar la demanda de asistencia a eventos deportivos, así como promover formatos de entretenimiento alternativos del deporte.

3.1– Revolución de los datos en el deporte

En las últimas dos décadas, la inteligencia artificial ha transformado la forma en que se consumen y analizan los deportes. El papel de la IA en la mejora de la toma de decisiones y la previsión en los deportes, entre muchas otras ventajas, se está expandiendo rápidamente y ganando más atención tanto en el sector académico como en la industria. No obstante, para muchas audiencias deportivas, profesionales y responsables políticos, que no son particularmente expertos en IA, la conexión entre la inteligencia artificial y los deportes sigue siendo difusa.

Del mismo modo, para muchos, las motivaciones para adoptar un paradigma de aprendizaje automático en el análisis deportivo aún son débiles o poco claras. En este documento

de perspectiva, se presenta una visión general de alto nivel, no técnica, del paradigma del aprendizaje automático que motiva su potencial para mejorar el análisis deportivo (rendimiento y negocios). Se Proporciona un resumen de alguna literatura de investigación relevante sobre las áreas en las que la inteligencia artificial y el aprendizaje automático se han aplicado a la industria del deporte y en la investigación deportiva. Finalmente, se presentan algunos escenarios hipotéticos de cómo la IA y el ML podrían dar forma al futuro de los deportes.

Es por esto que las organizaciones deportivas profesionales alguna vez vieron que los datos y el análisis tenían el potencial de ofrecer una ventaja informativa sobre la competencia. Hoy en día, la ciencia de datos de análisis deportivo es una apuesta de mesa. Estas organizaciones deben ir más allá del simple uso de datos para tomar decisiones y ejecutar nuevas ideas más rápido que la competencia, permitiendo a su gente tomar la mejor acción en el momento adecuado.

Fue el caso de Moneyball, que plantea el caso de éxito del equipo de béisbol de Grandes Ligas "Oakland Athletics", que el uso de estadísticas de juego en el juego se centró como un medio para armar un equipo excepcional. A pesar del presupuesto relativamente pequeño de los Atléticos de Oakland, la adopción de un enfoque riguroso basado en datos para armar un nuevo equipo llevó a los playoffs en el año 2002 [16].

Una revisión económica de la famosa hipótesis planteada en "Moneyball" arrojó luz sobre una interesante dinámica en el mundo del béisbol en ese momento. La investigación indicó que los salarios de los bateadores de béisbol no estaban necesariamente vinculados a su capacidad para contribuir al éxito del equipo en términos de victorias en los juegos. Esta revelación fue como una mina de oro para los Oakland Athletics, ya que les permitió obtener una ventaja competitiva distinta al identificar y capitalizar esta brecha de información en el mercado de jugadores.

Casi dos décadas después de que los principios de "Moneyball" y la era de la Sociedad para la Investigación del Béisbol Americano, estas metodologías han dejado una marca indeleble en la forma en que los equipos de las Grandes Ligas de Béisbol abordan el juego. Los científicos del deporte dedicados a recopilar y analizar datos de juego para responder preguntas clave relacionadas con el rendimiento del equipo, se han vuelto elementos comunes en las estructuras de los equipos de la MLB.

Esta evolución es un testimonio del poder transformador de la analítica de datos en el deporte, y cómo la combinación de estadísticas avanzadas y el enfoque científico ha redefinido la manera en que se construyen y gestionan los equipos de béisbol, y cómo se toman decisiones estratégicas en esta emocionante industria.

El aumento continuo y exponencial de la potencia de procesamiento de la computadora ha acelerado aún más la capacidad de analizar "big data" y, de hecho, las computadoras se están haciendo cargo cada vez más del análisis más profundo de los conjuntos de datos, a través de la inteligencia artificial (IA). Del mismo modo, el aumento en la recopilación de datos de alta calidad son ingredientes clave para el aumento en la precisión y amplitud de los análisis que se observó en la Grandes Ligas de Béisbol en los últimos años.

Por lo tanto, la adopción de la IA y el modelado estadístico en los deportes se ha vuelto más prominente en los últimos años, ya que las nuevas tecnologías y las aplicaciones de investigación están afectando a los deportes profesionales en varios niveles de sofisticación. La amplia aplicabilidad de los algoritmos de aprendizaje automático, combinada con el aumento de la potencia de procesamiento informático, así como el acceso a más y nuevas fuentes de datos en los últimos años, ha hecho que las organizaciones deportivas estén hambrientas de nuevas aplicaciones y estrategias.

El objetivo primordial sigue siendo hacerlos más competitivos dentro y fuera del campo, en rendimiento deportivo y comercial. Los beneficios de aprovechar el poder de la IA pueden, en ese sentido, tomar diferentes formas, desde la optimización de la toma de decisiones comerciales o técnicas hasta la mejora del rendimiento de los atletas / equipos, pero también el aumento de la demanda de asistencia a eventos deportivos, así como la promoción de formatos de entretenimiento alternativos del deporte.

Es por esto que en el 2020 la tecnología deportiva fue valorada en más de 11.700 millones de dólares, lo que da una muestra de su magnitud y la firma Vector ITC realizó un informe denominado "Digital Transformation in Sports", en él analiza las cuatro bases de la transformación digital en el terreno del deporte.

Los clubes están invirtiendo más que nunca en herramientas digitales relacionadas con el análisis del comportamiento de los usuarios y los deportistas, que permite anticipar sus necesidades dentro y fuera del estadio a través de Inteligencia Artificial, analítica de datos y Machine Learning, y seguir la evolución de KPIs enfocados en la relación con el cliente.

Los siguientes son algunos de los focos que han estado recibiendo cambios gracias a la aplicación del machine learning:

- **Fan Engagement,** ha mejorado significativamente gracias al Machine Learning. En la actualidad, las aplicaciones y plataformas de medios deportivos utilizan algoritmos para rastrear y analizar el comportamiento de los aficionados en línea. Esto permite la personalización de la experiencia del usuario, donde los aficionados reciben contenido relevante, como noticias, resúmenes de partidos y estadísticas de sus equipos y jugadores favoritos. Además, se utilizan análisis de sentimientos en redes sociales para comprender las opiniones de los aficionados y ajustar las estrategias de marketing y compromiso en tiempo real.
- **Optimización del Estadio,** en términos de energía, se emplean sistemas de gestión de edificios basados en Machine Learning para controlar la climatización y la iluminación según la ocupación, lo que reduce los costos energéticos. La seguridad se mejora mediante el análisis de video y la detección de comportamientos sospechosos en tiempo real, lo que garantiza un ambiente seguro para los aficionados. La experiencia del espectador se enriquece con aplicaciones de realidad aumentada y asistentes virtuales que proporcionan información y entretenimiento.
- **Excelencia Operativa,** lográndose mediante la eficiencia en la gestión. Las ligas y equipos deportivos utilizan el Machine Learning para optimizar la programación de eventos y partidos, maximizando la asistencia y la audiencia televisiva. También se

gestionan inventarios de productos y se optimiza la distribución, asegurando que los aficionados tengan acceso a alimentos, bebidas y mercancía de manera eficiente. El Machine Learning también se utiliza para predecir la demanda de entradas, lo que permite una planificación más precisa.

- **Rendimiento del Equipo,** esto se ha beneficiado enormemente del Machine Learning. Los equipos y entrenadores utilizan datos recopilados durante los partidos y entrenamientos para analizar el rendimiento individual y colectivo. Los algoritmos de Machine Learning identifican patrones en el juego, evalúan estadísticas clave y sugieren estrategias tácticas óptimas. Además, se emplean dispositivos wearables y sensores para monitorear la salud de los atletas, predecir lesiones y optimizar los planes de entrenamiento.

Estas prácticas 4 están siendo implementadas activamente en el fútbol, lo que ha llevado a mejoras significativas en la participación de los aficionados, la eficiencia operativa, la experiencia del estadio y el rendimiento del equipo. A medida que el Machine Learning continúa avanzando, se espera que estas aplicaciones se vuelvan aún más sofisticadas y efectivas.

Tal como es el caso del Sevilla F.C. que cuenta con un equipo de analistas que forman parte esencial de la estructura del club, y sus funciones son diversas y especializadas. Estos profesionales no solo se encargan de recopilar y analizar datos estadísticos de los partidos y entrenamientos, sino que también desempeñan un papel clave en la toma de decisiones tácticas y estratégicas.

En términos de análisis de rendimiento, estos analistas evalúan el desempeño individual de los jugadores en cada partido. Esto incluye el seguimiento de estadísticas detalladas, como la precisión de pases, los tiros a puerta y las intercepciones. Esta información es esencial para evaluar la contribución de cada jugador al equipo y para tomar decisiones sobre la alineación y las sustituciones.

Además del análisis de rendimiento, el equipo de analistas trabaja en estrecha colaboración con el cuerpo técnico para analizar la táctica del equipo y la de los oponentes. Utilizan datos para identificar patrones tácticos, puntos fuertes y debilidades, lo que influye directamente en la estrategia de juego y las decisiones tácticas tomadas durante los partidos.

Ellos tambien se dedican a la predicción y el modelado utilizando técnicas avanzadas de Machine Learning. Esto les permite predecir resultados de partidos futuros, evaluar el impacto de ciertos jugadores en el rendimiento general del equipo y realizar análisis más profundos de los datos disponibles.

Otro aspecto importante de su trabajo es el monitoreo de la salud de los jugadores. Utilizan datos de rendimiento físico y médico para identificar signos tempranos de fatiga o lesiones potenciales, lo que ayuda a prevenir problemas de salud y a optimizar los planes de entrenamiento.

3.2– Beneficios y desafíos

La ciencia de datos se está convirtiendo en una herramienta cada vez más importante en el mundo del deporte. Los equipos y atletas están utilizando datos para mejorar su rendimiento y tomar decisiones más informadas. En lo adelante se explora cómo se está aplicando la ciencia de datos al deporte y cuáles son sus beneficios y desafíos.

Los sistemas de aprendizaje automático han transformado la forma en que se aborda el análisis de datos en tiempo real en el fútbol y otros deportes de equipo. Estos sistemas tienen la capacidad de monitorear y analizar constantemente a todos los jugadores en el campo, así como la ubicación de la pelota, proporcionando un flujo constante de información valiosa para la toma de decisiones en el deporte.

Una de las aplicaciones más destacadas es la capacidad de evaluar la actividad de los jugadores mediante el cálculo de métricas esenciales, como la posesión del balón. Esto no solo permite un análisis en tiempo real, sino que también facilita una comprensión más profunda de cómo se desarrolla el juego y cómo interactúan los jugadores en el campo.

El poder predictivo del aprendizaje automático también se destaca, ya que puede predecir valores clave, como la posesión del balón, lo que es esencial para la toma de decisiones estratégicas. Por ejemplo, puede proporcionar información sobre cuándo aumentar la presión sobre el equipo contrario, cuándo adoptar una estrategia defensiva o cuándo buscar oportunidades de contraataque.

Otra aplicación valiosa es la capacidad de integrar datos del partido con información sobre la condición física de los jugadores. Esto brinda una visión completa del rendimiento de los atletas y puede ayudar al personal médico a evaluar el estado de los jugadores en tiempo real y tomar decisiones informadas sobre su participación en el juego.

Lo que hace que estos avances sean aún más significativos es su potencial para ser adoptados y utilizados de manera continua por los gerentes de equipo y los entrenadores. Si se integran de manera efectiva en la toma de decisiones estratégicas y tácticas, estos resultados pueden llevar a un fútbol más dinámico y efectivo. Estrategias como la contrapresión, que aprovecha la información en tiempo real para presionar al equipo contrario, o el contraataque, que se basa en la capacidad de predecir y aprovechar oportunidades, son ejemplos de cómo el aprendizaje automático está cambiando fundamentalmente la dinámica del juego [17].

La aplicación del aprendizaje automático en el ámbito del fútbol presenta un emocionante potencial para mejorar el rendimiento y la toma de decisiones. Sin embargo, esta innovación no está exenta de desafíos significativos.

Es fundamental asegurarse de la calidad de los datos. Los sistemas de aprendizaje automático dependen de datos precisos y completos para ofrecer resultados confiables. Errores o datos incompletos pueden llevar a predicciones erróneas y decisiones incorrectas. Por lo tanto, garantizar la calidad de los datos, desde la recopilación hasta el almacenamiento, es un desafío constante que debe abordarse meticulosamente.

Así como la implementación de sistemas de aprendizaje automático puede ser costosa, especialmente para equipos y organizaciones con recursos limitados. La inversión en tecnología y capacitación debe justificarse mediante mejoras demostrables en el rendimiento y la eficiencia. Encontrar un equilibrio entre los costos y los beneficios es un desafío clave en la adopción de esta tecnología.

El uso de datos en el deporte plantea preocupaciones importantes sobre la privacidad y la ética. La recopilación y el análisis de datos de jugadores y aficionados pueden invadir la privacidad de las personas, y es esencial respetar sus derechos. Además, se deben cumplir las regulaciones de protección de datos y establecer políticas sólidas para garantizar un uso ético de la información.

También la interpretación de los resultados del aprendizaje automático puede ser compleja y desafiante. Los entrenadores y el personal técnico deben comprender cómo utilizar la información proporcionada por los sistemas de manera efectiva en la toma de decisiones tácticas y estratégicas. La capacidad de traducir datos en acciones concretas es esencial para aprovechar al máximo el potencial de esta tecnología.

Y por último la evaluación de Impacto a largo plazo en el rendimiento del equipo puede ser un desafío complejo. Determinar qué métricas y resultados son más importantes y cómo medirlos de manera efectiva es esencial. Esta evaluación constante es necesaria para justificar inversiones continuas en tecnología y recursos.

3.3– Aplicaciones actuales

La ciencia de datos se está convirtiendo en una herramienta cada vez más importante en el mundo del deporte, de acuerdo a las investigaciones del machine learning y el fútbol, se puede constatar que existen numerosas aplicaciones con resultados positivos en el fútbol. Estas innovaciones están revolucionando cada aspecto, desde el análisis del rendimiento individual de los jugadores hasta la toma de decisiones estratégicas por parte de los entrenadores, y la entrega de experiencias más enriquecedoras para los aficionados. Hasta el momento se puede determinar que el machine learning se ha convertido en un pilar fundamental para la evolución y el éxito continuo del fútbol en la actualidad [18].

3.3.1. Análisis y mejora del desempeño

El uso del machine learning en el análisis del rendimiento ahora ayuda a los entrenadores a analizar el rendimiento de los jugadores en tiempo real e identificar áreas de mejora, al tiempo que proporciona información sobre el rendimiento individual y del equipo para un análisis táctico más profundo. Esta información también ayuda a identificar fortalezas y debilidades y a desarrollar estrategias de juego para optimizar los resultados de rendimiento.

Realizándose el **análisis del rendimiento individual** de los jugadores en cada partido, examinan datos como la distancia recorrida, la velocidad, los disparos a puerta, los pases completados y muchas otras métricas clave. Esto ayuda a los entrenadores y a los propios jugadores a identificar áreas de mejora en su juego, como la resistencia, la precisión en los tiros, la toma de decisiones y la capacidad de recuperación.

Además de la **evaluación objetiva del rendimiento de los jugadores**, lo que elimina la posibilidad de sesgos humanos. Esto es especialmente importante en la toma de decisiones relacionadas con la alineación del equipo y los cambios tácticos.

El machine learning puede generar recomendaciones inteligentes para el desarrollo de habilidades individuales basándose en los datos de rendimiento. Esto podría incluir ejercicios de entrenamiento específicos o consejos tácticos para mejorar el juego en ciertas posiciones.

Se han creado modelos de machine learning que pueden utilizar datos biométricos y de condición física de los jugadores para personalizar programas de entrenamiento. Esto significa que cada jugador puede tener un plan de entrenamiento específico que aborde sus debilidades y potencie sus fortalezas, lo que ayuda a maximizar el rendimiento individual.

Y por ultimo, **el análisis de Video Avanzado**, donde se realizan análisis de partidos para identificar patrones en el partido de un jugador. Esto incluye la forma en que se mueve, su posición en el campo y su interacción con otros jugadores. Siendo esta información es valiosa para ajustar la estrategia y tácticas individuales.

3.3.2. Predicción de resultados

La predicción de partidos de fútbol ha experimentado una evolución significativa gracias al machine learning. Esta tecnología ha permitido a los entusiastas del fútbol obtener pronósticos más precisos y detallados sobre los resultados de los juegos.

Los modelos de machine learning pueden analizan una amplia variedad de datos, que incluyen estadísticas de equipos y jugadores, historiales de enfrentamientos previos, condiciones meteorológicas y otros factores relevantes. Estos datos se utilizan para entrenar algoritmos que pueden predecir los resultados de los partidos con una precisión sorprendente.

Uno de los beneficios clave de la predicción de partidos basada en machine learning es su capacidad para considerar una gran cantidad de variables. Los modelos pueden tener en cuenta las tácticas de los equipos, las lesiones de los jugadores, las tendencias de rendimiento y muchos otros factores que pueden influir en el resultado de un partido.

Además, estos modelos no solo predicen el resultado final, sino que también pueden estimar la probabilidad de que ocurra un resultado específico, como un empate o un margen de victoria particular. Esto proporciona a los apostadores y fanáticos una visión más completa de lo que podría suceder en el campo.

3.3.3. Predicción de lesiones

La predicción de lesiones en jugadores es un aspecto crítico en el mundo del deporte, ya que muchas lesiones, como las distensiones musculares, pueden tener un impacto significativo en el rendimiento de un jugador y en el éxito del equipo. Detectar estas lesiones de manera temprana o predecir la probabilidad de que ocurran es esencial para la seguridad de los jugadores y la competitividad del equipo [19].

En un enfoque de aprendizaje supervisado, se recopilan datos detallados sobre los jugadores de temporadas anteriores. Estos datos incluyen información como el número de partidos jugados, el tiempo total en el campo, la edad de los jugadores, la distancia recorrida, si realizaron un calentamiento adecuado y cuántas veces fueron tacleados por otros jugadores. Sin embargo, el aspecto más crucial es si los jugadores sufrieron lesiones y se perdieron partidos posteriores.

La distinción clave en el aprendizaje supervisado es que se conoce el resultado a partir de datos históricos recopilados en temporadas anteriores. Estos datos se utilizan para entrenar un algoritmo de aprendizaje automático, que tiene como objetivo identificar los patrones o combinaciones de factores que aumentaron la probabilidad de lesiones. El modelo aprende de estos patrones y generalmente asigna una probabilidad de lesión basada en ellos.

Una vez que el modelo ha aprendido estos patrones, se pone a prueba con datos nuevos y no vistos para evaluar su capacidad para predecir lesiones con precisión. Si la precisión del modelo no cumple con los estándares requeridos, se ajusta y se entrena con parámetros ligeramente diferentes hasta que alcance el nivel de precisión deseado. Es importante destacar que este enfoque puede implementarse utilizando diversos algoritmos de aprendizaje automático, como redes neuronales, árboles de decisión y modelos de regresión, según lo necesario para los datos y los resultados deseados.

3.3.4. Estudio de los mercados deportivos

Dado que la mayoría de las organizaciones deportivas mantienen un seguimiento de los datos históricos relacionados con los patrocinadores que asisten a sus eventos deportivos, registrando una amplia gama de características como género, código postal, edad, nacionalidad, nivel educativo, ingresos, estado civil, entre otros, surge una pregunta natural de interés, eso con la intención de determinar si diferentes segmentos de clientes comprarán diferentes categorías de boletos por precio, duración o clase.

Mediante el empleo de análisis avanzados de datos, los algoritmos de Machine Learning tienen la capacidad de examinar vastos conjuntos de datos recopilados de diversas fuentes con el fin de identificar patrones, preferencias y tendencias. Este nivel de análisis profundo habilita a los especialistas en marketing deportivo para comprender con precisión y en tiempo real la interacción de la audiencia con el contenido y las campañas.

Mediante la identificación de patrones de comportamiento específicos se obtiene información valiosa sobre cuál contenido resuena de forma más efectiva en la audiencia y qué estrategias resultan más eficaces.

Estas percepciones permiten a los especialistas en marketing deportivo ajustar sus campañas de manera más efectiva, personalizando los mensajes y las estrategias para satisfacer las cambiantes expectativas de la audiencia. Como resultado, se fortalece la conexión entre el público y el deporte, ya que se toman decisiones de marketing más informadas y efectivas.

Adicionalmente, el Machine Learning es capaz de analizar rápidamente una gran cantidad de contenido, que puede variar desde artículos hasta vídeos y publicaciones en redes sociales, aprovechando los datos que reflejan el comportamiento y las preferencias de los usuarios.

Lo que diferencia al Machine Learning es que va más allá de simples recomendaciones, ya que comprende los intereses individuales de cada seguidor.

Al examinar patrones de comportamiento, preferencias previas y la interacción pasada con el contenido, el Machine Learning puede ofrecer recomendaciones altamente relevantes y atractivas para cada aficionado. Esto se traduce en la creación de experiencias personalizadas que se ajustan a los gustos específicos de cada persona, fomentando así una conexión más profunda con el equipo y el deporte en su conjunto. Al proporcionar una experiencia de seguidor enriquecida y personalizada, el Machine Learning refuerza la lealtad de los aficionados y aumenta su compromiso tanto con el deporte como con el equipo.

Por último, otros algoritmos de IA se diseñan con el propósito de segmentar los datos disponibles, de manera que cada punto de datos, como las ventas históricas de boletos, se clasifique en un grupo o clase similar a otros puntos de datos, basándose en las características registradas. Luego, estos algoritmos emplean métricas de similitud o distancia para clasificar a los usuarios en categorías de boletos que podrían adquirir, basándose en las tendencias observadas en los datos históricos.

3.3.5. Selección de jugadores

Los equipos deportivos en la actualidad destinan sumas significativas de recursos financieros para adquirir jugadores que puedan mejorar su rendimiento y competir casos de samente. Sin embargo, evaluar y seleccionar a los jugadores adecuados es un proceso enormemente complejo, ya que implica analizar una multitud de variables y parámetros que influyen en el rendimiento de un atleta durante el curso de un partido. La aplicación de métricas cuantitativas en tiempo real para la evaluación de jugadores individuales resulta un desafío insuperable para el análisis humano, dado que la velocidad y precisión requeridas son prácticamente inalcanzables.

Es en este contexto que el Machine Learning y la Inteligencia Artificial han emergido como herramientas esenciales para la evaluación de jugadores en el mundo del deporte. Estas soluciones avanzadas hacen uso de los datos recopilados a partir de la cobertura de video y los dispositivos instalados en los estadios para llevar a cabo un análisis minucioso de los jugadores en tiempo real.

Mediante el Machine Learning, es posible medir una amplia gama de parámetros relacionados con el desempeño de los jugadores. Los equipos pueden, por ejemplo, comparar la habilidad de un jugador con la de un experto que desean incorporar a su equipo, permitiendo tomar decisiones más fundamentadas en el proceso de reclutamiento. Además, estas soluciones de IA pueden profundizar en la fisiología de cada jugador, brindando información invaluable sobre cómo funcionan bajo estrés y en situaciones específicas del juego.

El proceso de reclutamiento en la industria deportiva es singular en comparación con otros sectores, ya que no solo implica el análisis de estadísticas históricas de un deportista, sino también la evaluación en tiempo real de su rendimiento actual. Esta tarea no sería posible sin la asistencia de la Inteligencia Artificial, que es capaz de detectar y clasificar diferencias mínimas entre jugadores y seleccionar a los candidatos más destacados. En esencia, el Machine Learning y la IA han redefinido la forma en que los equipos deportivos identifican y adquieren talento, permitiéndoles tomar decisiones más precisas y estratégicas en el proceso de reclutamiento, lo que finalmente contribuye a mejorar el rendimiento y el éxito del equipo en competencias futuras.

3.3.6. Mejores estrategias deportivas

El análisis de los oponentes es otro campo donde el Machine Learning brilla con luz propia. Estos algoritmos pueden procesar grandes cantidades de datos de partidos anteriores de los rivales, identificando patrones y tendencias en su estilo de juego. Esto permite a los equipos anticipar las jugadas más probables de los oponentes y preparar estrategias defensivas adecuadas. Por ejemplo, si se sabe que un equipo rival tiende a centrar el balón en jugadas a balón parado, se pueden ajustar las tácticas defensivas y la alineación para contrarrestar esta amenaza específica.

Durante los partidos en tiempo real, el Machine Learning continúa siendo un aliado valioso. Los sistemas pueden proporcionar análisis instantáneos, ayudando a los entrenadores a tomar decisiones cruciales, como realizar sustituciones o ajustar tácticas sobre la marcha. Por ejemplo, si un análisis en tiempo real revela que el equipo está perdiendo la posesión en el centro del campo, el entrenador puede optar por reforzar el mediocampo para recuperar el control del juego. Sin embargo, el impacto del Machine Learning va más allá de

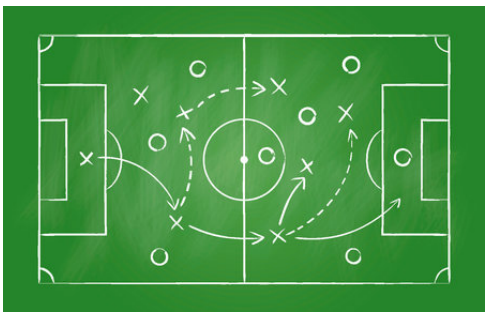


Figura 3.1: Estrategia de movimientos en campo.

la toma de decisiones tácticas durante los partidos. Estos sistemas también ofrecen una visión profunda del rendimiento general del equipo y los patrones de juego a lo largo de una temporada. Los análisis posteriores a los partidos proporcionan estadísticas detalladas, desde la eficacia en los pases hasta las áreas del campo donde un equipo tuvo más éxito. Esta información es fundamental para refinar las estrategias de juego a largo plazo, identificar áreas de mejora y destacar las fortalezas que deben explotarse.

En última instancia, el Machine Learning no solo se trata de mejorar las estrategias de juego en el presente, sino también de establecer las bases para un éxito sostenible en el futuro. Los equipos utilizan estos sistemas para evaluar el rendimiento de sus jugadores y

tomar decisiones fundamentales, como la contratación de nuevos talentos o la inversión en el desarrollo de jóvenes promesas. Al comprender mejor los patrones y tendencias a lo largo del tiempo, los equipos pueden construir plantillas sólidas y rentables que se mantengan en la cima del juego durante años.

3.4– Casos de éxito

En esta sección, se exploran algunos casos destacados en los que el machine learning ha transformado el futbol y ha impulsado a equipos y jugadores hacia nuevos niveles.

3.4.1. Sports Performance Platform

La Sports Performance Platform de Microsoft es una solución de análisis deportivo personalizada diseñada para brindar a atletas y equipos deportivos la capacidad de tomar decisiones más informadas y basadas en datos. Esta plataforma ha sido utilizada por equipos de renombre como Seattle Reign FC, Real Sociedad, Sport Lisboa e Benfica y Cricket Australia, destacándose por su capacidad para rastrear y mejorar el rendimiento de los jugadores.

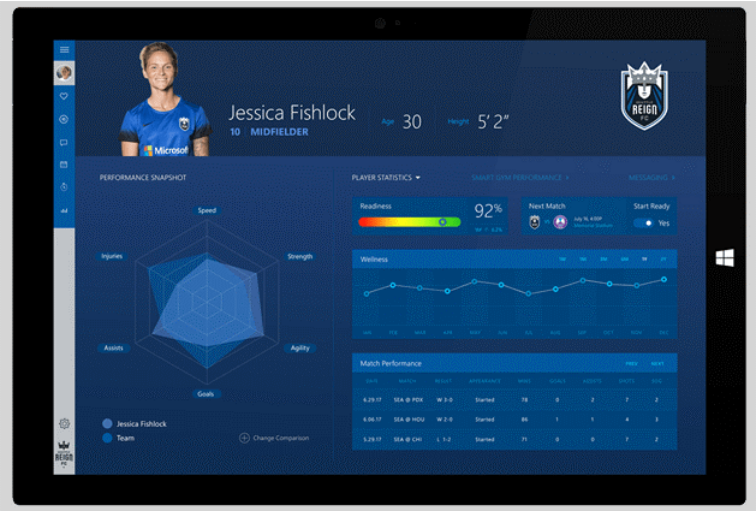


Figura 3.2: Visualización en Sports Performance plataform.

Microsoft
Sports Performance Platform

Figura 3.3: Logo de Sports Performance plataform.

El objetivo de esta tecnología es democratizar el acceso a las herramientas de análisis y modelado de resultados predictivos en el mundo del deporte, abriéndolas a todos los niveles, desde el profesional hasta el académico y los deportistas individuales. Desarrollada sobre dispositivos como Power BI, Microsoft Azure y Microsoft Surface, Sports Performance Platform aprovecha los avances en la computación en la nube, la agregación de datos, el aprendizaje automático y el análisis predictivo.

Los entrenadores pueden utilizar esta plataforma para crear una visión integral del rendimiento, la recuperación y la preparación de un atleta. Esto les permite ejecutar modelos predictivos para prevenir lesiones, tomar decisiones de último minuto sobre la disponibilidad de un jugador en el día del partido y diseñar regímenes de entrenamiento específicos para mantener a los atletas en su mejor forma. Microsoft ha colaborado con socios como Akvelon, Fair Play y POP para ayudar a los clientes a implementar y personalizar Sports Performance Platform según sus necesidades específicas.

Un ejemplo concreto de su aplicación es el Seattle Reign FC, uno de los primeros equipos en adoptar esta tecnología. Aquí, los datos de GPS de Catapult y la plataforma de monitoreo Fit For 90 se utilizan para rastrear una variedad de métricas, desde la frecuencia cardíaca hasta la velocidad y la aceleración de los jugadores. Sports Performance Platform agrega todos estos datos dispares en una vista centralizada [20].

Después de cada partido, el director de rendimiento, puede revisar estos datos para evaluar el rendimiento de los jugadores. Esto incluye información sobre cuánto corrieron durante el partido, lo que podría indicar un mayor riesgo de dolor muscular, así como la posible fatiga cardiovascular debido a condiciones climáticas extremas. Esta información permite al equipo tomar decisiones informadas sobre la recuperación y el entrenamiento de los jugadores.

Además, la plataforma permite a los entrenadores y al personal de rendimiento del Reign realizar un seguimiento del progreso de los atletas a lo largo de múltiples temporadas, identificando a jugadores con las habilidades necesarias para destacar en el nivel profesional. También se planea extender su uso a la Reign Academy, una iniciativa de desarrollo de jóvenes talentos en el fútbol. Sports Performance Platform está transformando la forma en que los equipos de fútbol gestionan y optimizan el rendimiento de sus jugadores.

3.4.2. Principios de Johan Cruyff

En junio de 2019, el Barça Innovation Hub presentó su investigación en la conferencia mundial MIT Sloan Sports Analytics. Su estudio se enfocó en desarrollar un modelo matemático innovador que evalúa la calidad de las decisiones tomadas por los jugadores de fútbol en función de la posición de sus compañeros y rivales en cada momento del partido [21].

Este modelo se basa en un concepto llamado “Expected Possession Value”, o Valor de Posesión Esperada, que se centra en determinar la calidad de la disposición de los jugadores en el campo en tiempo real y cómo esto influye en la probabilidad de que una posesión termine en gol. Lo que hace único a este enfoque es que considera el contexto en el que ocurren las acciones, en contraposición a las estadísticas tradicionales que se centran en acciones individuales aisladas, como pases exitosos o posesión del balón.

Para construir este modelo, el equipo de investigadores analizó la posición de los jugadores en videos de partidos del FC Barcelona durante dos temporadas. Utilizando algoritmos de Machine Learning, pudieron calcular el impacto, segundo a segundo, de la disposición de los jugadores en la probabilidad de que se produzca un gol.

Esta herramienta tiene un gran potencial y ha sido desarrollada en colaboración con entrenadores, analistas de juego y científicos de datos del club. Puede aplicarse en diversas situaciones prácticas, como evaluar el riesgo y el beneficio de los pases, identificar ventajas posicionales según el contexto y ofrecer opciones alternativas para cada jugador en tiempo real. Este ejemplo ilustra cómo la inteligencia artificial está transformando el fútbol, con un enfoque técnico orientado al usuario final, en este caso, los entrenadores.

3.4.3. Probabilidad de gol por reconocimiento de video LaLiga

Otro caso destacado es el de LaLiga, que está estableciendo un precedente en las ligas europeas, al introducir un innovador modelo de probabilidad de gol en sus transmisiones de partidos a partir de la jornada 22 en adelante. Esta métrica revolucionaria se ha desarrollado en colaboración con Microsoft y se basa en tecnologías avanzadas de inteligencia artificial y aprendizaje automático.

La métrica de probabilidad de goles de LaLiga y Microsoft evalúa parámetros específicos mediante un modelo matemático que determina cómo cada uno de estos factores influye en la probabilidad de que se anote un gol. Durante un partido en tiempo real, se recopilan datos sobre la ubicación de los jugadores gracias a las cámaras de seguimiento óptico instaladas en cada estadio de la liga. A partir de estos datos, se calculan los parámetros y el modelo genera la Probabilidad de Gol.

Figura 3.4: Probabilidad de Gol en LaLiga.

Esta métrica mejorada se basa en una variedad de variables, como la línea de visión del jugador, que considera las posiciones de los jugadores rivales que podrían obstaculizar la visión del jugador hacia la portería. Esto tiene un impacto significativo en la dificultad y, por lo tanto, en la probabilidad de que se aproveche la oportunidad. También se tienen en cuenta otros factores, como la distancia entre el balón y el portero, la distancia entre el balón y la portería, y el ángulo con el defensor más cercano. Estos datos se traducen en una estadística final que se muestra en pantalla. Cuanto menor sea la probabilidad de marcar un gol para un tiro que realmente entra, más eficiente habrá sido el disparo en comparación con lo que cabría esperar del jugador.

Esta métrica se ha desarrollado a partir de un modelo interno por parte de los analistas de fútbol de Mediacoach de LaLiga, en colaboración con el equipo de Business Intelligence & Analytics de LaLiga Tech, la filial tecnológica de la organización. La característica más notable y diferenciadora de esta métrica es su capacidad para incorporar, casi en tiempo real, las estadísticas históricas del jugador durante la transmisión televisiva en las segundas repeticiones de los goles. Esta innovación ha sido posible gracias a la tecnología de Microsoft Azure [24].

3.5– Herramientas y tecnologías

3.5.1. Herramientas y software especializado

Existen muchas herramientas y entornos que son ampliamente utilizados para aplicar el machine learning. Estas herramientas ayudan a los científicos de datos, ingenieros de datos y desarrolladores a construir, entrenar y desplegar modelos de machine learning. Algunas de las herramientas más importantes y populares se describen a continuación.

Python:

Es un lenguaje de programación versátil y de código abierto que se ha convertido en el estándar de facto en la comunidad de aprendizaje automático. Ofrece una amplia gama de bibliotecas y frameworks especializados, lo que facilita el desarrollo de modelos y la manipulación de datos. Algunos de los paquetes más populares para el aprendizaje automático en Python incluyen Pandas para manipulación de datos, NumPy para cálculos numéricos y Matplotlib/Seaborn para visualización.

Figura 3.5: Logo del lenguaje Python.

R:

R es un lenguaje especializado en estadísticas y análisis de datos ampliamente utilizado en la investigación académica y preferido por estadísticos y científicos de datos debido a su flexibilidad y potencia estadística. Facilita la exploración de datos, modelado estadístico y visualización, ofreciendo una amplia gama de paquetes y bibliotecas.

Figura 3.6: Logo del lenguaje R.

TensorFlow:

Es uno de los framework de deep learning más influyentes. Ofrece una arquitectura flexible para construir y entrenar redes neuronales profundas. Además, TensorFlow ofrece TensorFlow.js para el aprendizaje automático en el navegador y TensorFlow Lite para aplicaciones móviles.

Figura 3.7: Logo del framework TesorFlow.

PyTorch:

Creado por Facebook, se ha vuelto muy popular en la investigación de aprendizaje profundo debido a su flexibilidad y flujo de trabajo más intuitivo. Su estructura dinámica permite la creación rápida de prototipos de modelos, y PyTorch ofrece herramientas como TorchScript para la implementación en producción.

Figura 3.8: Logo de librería PyTorch.

Scikit-Learn

Es una biblioteca de aprendizaje automático en Python que se centra en algoritmos tradicionales de machine learning. Ofrece una amplia gama de herramientas para tareas comunes en la ciencia de datos, como clasificación, regresión, clustering y reducción de dimensionalidad. Scikit-Learn es especialmente adecuada para proyectos donde no se necesitan modelos de deep learning, ya que se enfoca en algoritmos más tradicionales pero altamente efectivos.

Figura 3.9: Logo de librería Scikit-Learn.

Keras:

Es una API de alto nivel que funciona sobre varios backends de deep learning, incluido TensorFlow. Se destaca por su facilidad de uso y se utiliza para crear y entrenar rápidamente modelos de redes neuronales, especialmente por aquellos nuevos en el aprendizaje profundo.

Figura 3.10: Logo de Keras.

Jupyter Notebooks:

Son una herramienta esencial para la ciencia de datos y el aprendizaje automático. Permiten la creación de documentos interactivos que combinan código, visualizaciones y texto explicativo. Son ideales para la exploración de datos y la presentación de resultados.

Figura 3.11: Logo de Jypyter.

Apache Spark:

Apache Spark es un poderoso framework de procesamiento de datos en memoria que se ha vuelto esencial en el campo de la ciencia de datos y el aprendizaje automático. Este framework no solo acelera significativamente el procesamiento de datos, sino que también ofrece una biblioteca llamada MLlib que es fundamental para el aprendizaje automático distribuido. Lo que hace que Apache Spark sea especialmente valioso es su capacidad para lidiar con grandes volúmenes de datos y su habilidad para escalar horizontalmente, lo que significa que puede manejar conjuntos de datos masivos y tareas de procesamiento intensivas de manera eficiente.machine learning a gran escala.

Figura 3.12: Logo del framework Apache Spark.

Microsoft Azure ML:

Es una plataforma en la nube que ofrece herramientas para el desarrollo, entrenamiento y despliegue de modelos de aprendizaje automático. Facilita la colaboración en equipo y la implementación en la nube de Microsoft.

Figura 3.13: Logo de Azure Machine Learning.

Google Cloud AI Platform:

Esta plataforma de Google Cloud proporciona una amplia gama de servicios para el aprendizaje automático, desde la construcción de modelos hasta la implementación en la nube de Google.

Figura 3.14: Logo de plataforma Google cloud plataform.

Google Colab:

Es una herramienta importante en el campo del aprendizaje automático. Se basa en los Jupyter Notebooks y se ejecuta en la infraestructura de Google, lo que lo convierte en una opción popular para aquellos que desean aprovechar el poder de la nube de Google para ejecutar código de Python y realizar tareas de aprendizaje automático.

Figura 3.15: Logo de plataforma Google Colab.

3.5.2. Tecnologías en uso

En este apartado se explican las tecnologías clave en uso dentro del ámbito del fútbol que están impulsando avances en el análisis de datos, el rendimiento de los jugadores y la experiencia del espectador. Cada tecnología se presentará junto con la empresa que la proporciona y se destacarán ejemplos específicos de su implementación. Explicando una explicación general de las herramientas y sistemas que están transformando la forma en que se juega y se disfruta el fútbol en la era digital.

PlayerMaker

Mediante el uso del wearable, que se coloca de manera conveniente en los zapatos del jugador, PlayerMaker ofrece una plataforma excepcionalmente rica en análisis e información. Esta plataforma no solo registra el rendimiento en el campo, sino que también proporciona valiosos enlaces al sueño y la nutrición, características que rara vez se encuentran en sistemas de seguimiento de datos.

Su gran variedad de datos permite a los jugadores una adaptación precisa de sus cargas físicas en términos de descanso y recuperación. En otras palabras, PlayerMaker brinda la capacidad de crear las condiciones óptimas para que el talento natural de un jugador florezca de manera excepcional.

Figura 3.16: Logo de plataforma PlayerMaker y el dispositivo de rastreo.

La aplicación de PlayerMaker, diseñada de manera inteligente y de fácil navegación, complementa esta experiencia de seguimiento. Su utilización efectiva de infografías y gráficos permite una interpretación clara y concisa de los datos recopilados. Además, ofrece una característica única y emocionante: la posibilidad de observar las mejores marcas personales en cada métrica y compararlas con otros usuarios de PlayerMaker en todo el mundo, incluso con profesionales de renombre.

Esta característica no solo impulsa una competencia saludable, sino que también proporciona una perspectiva global del rendimiento, lo que resulta especialmente valioso para los equipos al permitirles comparar el desempeño de sus jugadores con los estándares internacionales. En resumen, PlayerMaker brinda un sistema completo e integral para el seguimiento y análisis del rendimiento en el fútbol, ofreciendo una ventaja competitiva significativa tanto para los jugadores como para los equipos.

Sports Performance Platform

Como se había detallado anteriorment en los casos de éxito, se habla que Sports Performance Platform de Microsoft es una solución tecnológica de vanguardia diseñada específicamente para equipos deportivos y organizaciones relacionadas con el deporte. Esta plataforma representa un salto significativo en la optimización del rendimiento deportivo y la toma de decisiones estratégicas. Lo que distingue a esta plataforma es su capacidad para recopilar datos en tiempo real de diversas fuentes, como sensores en el cuerpo de los atletas, dispositivos de seguimiento y cámaras. Esta capacidad permite un monitoreo constante y preciso del rendimiento de los deportistas durante entrenamientos y competencias, proporcionando información valiosa para entrenadores y jugadores.

Una de las características más destacadas de Sports Performance Platform es su capacidad de análisis avanzado. Los datos recopilados se someten a un riguroso proceso de análisis mediante algoritmos de machine learning e inteligencia artificial. Esto da como resultado información detallada sobre el rendimiento físico, táctico y técnico de los jugadores, así como sobre las tendencias y patrones emergentes en el deporte. La plataforma también ofrece herramientas de visualización de datos de vanguardia, como gráficos interactivos y tableros personalizables, lo que facilita la comprensión de la información y la toma de decisiones informadas.

La planificación y la estrategia son componentes fundamentales de Sports Performance Platform. Los equipos deportivos pueden utilizar esta plataforma para planificar estrategias de entrenamiento, tácticas de juego y gestión de equipos basadas en los datos recopilados. Además, la plataforma ayuda en la gestión de lesiones al proporcionar información valiosa sobre la salud y el estado físico de los atletas. La colaboración también se ve beneficiada, ya que la plataforma facilita la comunicación y el intercambio seguro de datos entre entrenadores, jugadores y personal de apoyo.

IBM Watson Discovery

IBM Watson Discovery es una plataforma de procesamiento de lenguaje natural (NLP) y búsqueda cognitiva desarrollada por IBM que juega un papel crucial en la transformación del análisis de datos en el fútbol.

Su funcionamiento se basa en técnicas avanzadas de Machine Learning, que le permiten procesar y analizar grandes volúmenes de información no estructurada relacionada con el deporte, incluyendo metraje de video, transcripciones de audio y texto. Su objetivo principal es extraer conocimientos significativos y valiosos de estos datos, que a menudo son difíciles de analizar manualmente, para brindar información relevante a entrenadores, jugadores, analistas y aficionados.

IBM Watson Discovery en el fútbol aporta numerosos beneficios revolucionarios a la industria deportiva. Esta plataforma de procesamiento de lenguaje natural y búsqueda cognitiva ha permitido una indexación eficiente de vastos volúmenes de contenido relacionado con el fútbol. Esto significa que ahora es más fácil que nunca buscar momentos clave en los partidos o analizar detenidamente el rendimiento de los jugadores.

Además, IBM Watson Discovery incorpora una funcionalidad de análisis de sentimiento que puede evaluar el tono y la percepción pública en noticias, comentarios y redes sociales. Esto proporciona información valiosa sobre cómo se percibe el juego y sus protagonistas en la esfera pública, lo que puede ser esencial para la gestión de la reputación de los equipos y jugadores.

Figura 3.17: Logo de plataforma IBM Warson Discovery.

El reconocimiento de patrones, impulsado por algoritmos de aprendizaje automático, es otra ventaja fundamental de esta tecnología. Ayuda a identificar patrones en el juego y el desempeño de los jugadores, lo que facilita la toma de decisiones estratégicas y tácticas en el fútbol.

Una característica significativa es la capacidad de IBM Watson Discovery para generar resúmenes inteligentes, lo que agiliza la creación de contenido destacado para los medios de comunicación y permite a los equipos revisar rápidamente los aspectos más relevantes de los partidos.

La búsqueda avanzada es otra ventaja crucial, ya que permite realizar búsquedas específicas en grandes conjuntos de datos de video y texto de manera rápida y efectiva, ahorrando tiempo y proporcionando información detallada sobre partidos anteriores y estadísticas.

Otras tecnologías en auge

Stats Perfom: es una empresa especializada en análisis y datos deportivos que ha surgido como resultado de la fusión de dos empresas prominentes en este campo, Stats y Perform. Esta empresa se ha convertido en una de las principales actores en la recopilación y análisis de datos deportivos, y su alcance abarca una amplia variedad de sectores en la industria deportiva.

Figura 3.18: Logo plataforma Stats Perform.

Uno de los aspectos clave de Stats Perform es su enfoque en el análisis predictivo. Utiliza datos recopilados de eventos deportivos en todo el mundo para proporcionar información valiosa que se utiliza en diversos campos dentro del deporte. Esto incluye el análisis del rendimiento de equipos profesionales, lo que ayuda a los entrenadores y gerentes a tomar decisiones informadas sobre estrategias y alineaciones.

Además, Stats Perform trabaja con medios digitales, lo que significa que sus datos se utilizan para mejorar la cobertura y presentación de eventos deportivos en plataformas de medios. Esto puede incluir estadísticas en tiempo real, visualizaciones de datos y análisis que enriquecen la experiencia de los espectadores. La empresa también tiene una presencia en el mundo de las apuestas deportivas y los deportes de fantasía. Proporciona datos y análisis que ayudan a los apostadores y a los aficionados de los deportes de fantasía a tomar decisiones fundamentadas.

Un aspecto interesante es el enfoque de Stats Perform en la inteligencia artificial y el aprendizaje automático. Estas tecnologías se utilizan para analizar grandes conjuntos de datos y descubrir patrones y tendencias que a menudo pasan desapercibidos para los analistas humanos.

3.6– Futuro del Machine Learning en el deporte

La Inteligencia Artificial (IA) continuará transformando el mundo de los deportes, y las implicaciones de esta transformación serán sorprendentes e innovadoras. En la actualidad, el Machine Learning ha alterado significativamente la forma en que comprendemos y abordamos las estrategias de juego y el análisis del rendimiento de los jugadores. Además ha modificado la perspectiva sobre la audiencia y los consumidores de deportes.

Figura 3.19: Machine Learning en campo de Fútbol.

Es evidente que se está ingresando en un territorio donde la IA se integra en todas las facetas de los deportes, desde la toma de decisiones tácticas hasta la creación de contenido deportivo personalizado. Esto plantea cuestiones éticas que deben ser cuidadosamente consideradas a medida que las máquinas desempeñan un papel más prominente en un ámbito deportivo tradicionalmente centrado en el ser humano y el talento natural de los atletas.

Si bien es poco probable que la IA reemplace completamente a entrenadores y expertos humanos, su capacidad para proporcionar información valiosa y ventajas competitivas tanto a los equipos como a los especialistas en marketing deportivo es innegable. Los datos en tiempo real sobre el comportamiento de la audiencia permiten ajustar campañas y estrategias de manera eficaz para satisfacer las cambiantes expectativas de los fanáticos.

Reemplazar a los árbitros con IA automatizada, si bien es técnicamente posible y cada vez más utilizado en algunos deportes debido a su precisión y eficiencia, plantea la pregunta crucial de si esto es lo que realmente desean los fanáticos [22].

Aunque la IA desempeñará un papel cada vez más significativo en los deportes, es esencial reconocer que la dirección y la intervención humanas seguirán siendo fundamentales tanto para el rendimiento deportivo de élite como para la toma de decisiones estratégicas en los negocios deportivos. La magia del deporte radica en su capacidad para unir lo humano y lo tecnológico, proporcionando una experiencia única que continúa emocionando a los fanáticos en todo el mundo.

Desde una perspectiva económica, confiar en algoritmos artificiales podría aumentar los ingresos de las organizaciones deportivas y los organizadores de eventos al permitirles aplicar estrategias eficientes de precios variables y dinámicos. Además, estos algoritmos pueden ayudar a construir plataformas de consumo más profundamente personalizadas.

La aplicación de diversos algoritmos de aprendizaje automático posibilita un marketing más eficaz al impulsar la personalización y mejorar las tasas de conversión en el proceso de ventas.

Sin embargo, el futuro de la privacidad de los datos plantea preguntas significativas. La utilización de algoritmos de IA para influir en el comportamiento humano y tomar decisiones basadas en datos conlleva desafíos éticos, estas cuestiones deben abordarse a medida que se avanza en la era de la IA en el deporte [23].

CAPÍTULO 4

Visión general de las pruebas de conceptos

4.1– Introducción y características del dominio

Como se ha observado en los casos de éxito anteriores, las tecnologías relacionadas con el análisis de datos y el machine learning han revolucionado la industria del fútbol al ofrecer la capacidad de medir el rendimiento de los jugadores, predecir resultados de partidos y proporcionar información detallada sobre aspectos específicos de sus actuaciones.

Las herramientas de Machine Learning con que se cuenta actualmente en el fútbol han demostrado ser invaluables para los equipos al momento de tomar decisiones estratégicas lo pueden hacer de forma informada.

Por esta razón, se ha optado por llevar a cabo una serie de pruebas de conceptos que involucran técnicas de análisis de datos y machine learning que han demostrado ser particularmente prometedoras en casos de éxito previos. Estas pruebas de conceptos se centran en dos áreas clave:

- **Análisis de Datos del Rendimiento de los Jugadores:** se centra en el análisis exhaustivo de los datos de rendimiento de los jugadores. Esto implica la recopilación y evaluación de métricas como el tiempo de posesión del balón por jugador/equipo, las distancias recorridas en el campo (en caso de disponibilidad de datos) y la efectividad de los pases. Estos datos proporcionarán información valiosa sobre el rendimiento individual y colectivo, permitiendo a los equipos identificar patrones y áreas de mejora.

- **Predicción mediante Machine Learning:** se enfoca en la aplicación de técnicas de machine learning para realizar predicciones significativas en el contexto del fútbol. Aunque la selección de técnicas específicas aún está por determinar, el objetivo es utilizar machine learning para predecir resultados de partidos, anticipar posibles lesiones y, si es viable, estimar goles esperados tanto a nivel individual como de equipos. Estas predicciones pueden ser valiosas para los equipos al tomar decisiones estratégicas y tácticas antes de los partidos.

4.2– Objetivos

El objectivo de este capítulo es el desarrollo de pruebas de conceptos, basada en trabajos previos, e implementando las funcionalidades probadas y exitosas en el análisis de datos y el machine learning aplicado al fútbol.

De igual forma, otras metas que se buscan alcanzar son:

- Profundizar en la comprensión del rendimiento de los jugadores en el fútbol, explorando estadísticamente los datos disponibles, para identificar patrones y entender cómo los jugadores contribuyen al éxito de sus equipos. El desafío de traducir estos números en estrategias y habilidades aplicables en el mundo real resulta sumamente fascinante.
- Emplear algoritmos y modelos para predecir resultados de partidos, anticipar lesiones y evaluar el desempeño de los jugadores resulta especialmente atractiva. A pesar de ser un principiante en el mundo del fútbol, una motivación es la idea de contribuir a la creación de tecnologías que puedan aportar un valor significativo. Además, esta área ofrece oportunidades continuas de investigación y desarrollo, lo que impulsa a seguir aprendiendo y perfeccionando las habilidades en el campo del machine learning.

4.3– Trabajos relacionados

Durante la investigación, se encontraron numerosos trabajos en línea que emplean el machine learning para realizar análisis y predicciones en el fútbol.

Uno de estos trabajos destacados es el realizado por Oscar Bartolomé Pato como parte de su Tesis de Maestría, titulado: **"Creación de un Modelo de Clasificación para Calcular la Métrica XG"** [25]. El objetivo principal de este proyecto es calcular la probabilidad de que un disparo se convierta en gol, basándose en una serie de características. Estas características incluyen elementos previos al disparo, como la posición del tiro, el ángulo con respecto a la línea de gol y la parte del cuerpo utilizada para rematar, entre otros.

Este estudio utilizó datos de 5 partidos para entrenamiento y posteriormente realizó predicciones. A través del uso de un modelo de regresión logística y XGBoost, logró alcanzar una precisión del 90 % en la predicción de si un disparo resultaría en gol o no.

Otro trabajo relevante se centra en la predicción del rendimiento de un jugador de fútbol. Fue realizado por Iñigo Gómez como parte de su Trabajo de Fin de Grado en la Universidad Autónoma de Madrid [26].

Este proyecto busca un enfoque analítico que no solo tome en cuenta el valor estadístico generado por un jugador, sino también el equipo en el que juega, sus características individuales, nivel y estilo de juego. El objetivo es proporcionar una herramienta para que los aficionados evalúen a los jugadores de manera más precisa. Además, se incorporan modelos de aprendizaje automático para evaluar si la inteligencia artificial puede predecir el rendimiento de un jugador en el próximo año.

En este trabajo, se implementaron modelos predictivos como ElasticNet, Random Forest y XGBoost, que arrojaron métricas MAE (Fórmula Error absoluto medio) con valores de 0.23 a 0.67 y RMSE (Error cuadrático medio) de 0.30 a 0.87.

Estos son solo algunos de los muchos trabajos de machine learning aplicados al mundo del fútbol. También se han investigado otros enfoques, como la predicción de resultados de partidos, la identificación del equipo ganador, las posibilidades de empate y la predicción de posibles lesiones de los jugadores.

Otro trabajo relacionado al análisis de datos es el de la estudiante de la univerdad de Valencia, María del Pilar Malagón Selma, que en su trabajo de fin de master con titulo: **"Machine Learning en el mundo del fútbol"** [27], analiza el rendimiento en el terreno de juego de atletas profesionales de las grandes ligas europeas de fútbol durante la temporada 2017-18 a partir de estrategias y técnicas de Machine Learning y Big Data como el análisis de componentes principales, análisis de clúster, gráficos de contribuciones y gráficos de radar. Se profundiza en las posiciones ideales de los futbolistas en el campo y se identifican las variables más influyentes en cada demarcación a partir del modelo Random Forest.

Esta investigación supone una propuesta metodológica novedosa en relación al análisis de datos en el fútbol gracias al uso de programas informáticos como R, Aspen ProMV y Tableau con la finalidad de facilitar el proceso de captación de talento en las direcciones deportivas de los clubes del fútbol profesional.

En definitiva, este trabajo de investigación ofrece un conjunto de herramientas basadas en el análisis de datos para mejorar la toma de decisiones y reducir la incertidumbre a la hora de acometer la contratación de un nuevo jugador en el fútbol de élite.

Esta investigación ofrece un conjunto de herramientas basadas en el análisis de datos para mejorar la toma de decisiones y reducir la incertidumbre en la contratación de nuevos jugadores en el fútbol de élite.

Los anteriores solamente son dos ejemplos entre numerosos trabajos de Machine Learning aplicados al mundo de futbol, por lo que también se pueden ver otros trabajos relacionados a la predicción de partidos, predicción del equipo ganador, la posibilidad de empate, o la predicción de posibles lesiones de los jugadores.

Ha sido precisamente la gran propularidad del Fútbol lo que ha permitido que tanto a nivel educativo como profesional y científico se lleven a cabo investigaciones con el fin de seguir creando herramientas que proporcionen buenas métricas a los equipos, y como fin último, que se brinde una buena experiencia a los fanáticos.

4.4– Descripción y Alcance de las pruebas

En el contexto de estas pruebas de conceptos, se busca desarrollar una comprensión más profunda y específica del rendimiento en el fútbol a través de la aplicación de técnicas de análisis de datos y machine learning. El objetivo primordial de esta iniciativa es explorar y evaluar cómo estas tecnologías pueden impactar positivamente en la toma de decisiones en el fútbol.

El alcance de estas pruebas abarca dos áreas principales:

- **Análisis de Datos de Jugadores**

  En esta fase, se llevará a cabo una exhaustiva recopilación y análisis de datos relacionados con el rendimiento de los jugadores. Esto incluye aspectos como el tiempo de posesión del balón por parte de jugadores y equipos, las distancias recorridas en el campo de juego (sujeto a la disponibilidad de los datos), y la efectividad de los pases de balón. El objetivo es desglosar y comprender cómo estos datos pueden revelar patrones y tendencias que influyen en el éxito de los jugadores y sus equipos, y permite una comprensión profunda del rendimiento de los jugadores y los equipos. Esto puede llevar a la identificación de áreas de mejora tanto a nivel individual como colectivo.

- **Aplicación de Técnicas de Machine Learning**

  En esta etapa, se emplearán algoritmos y modelos de machine learning para realizar predicciones y análisis más profundos. Se considerará la posibilidad de predecir resultados de partidos, los goles esperados de jugadores y equipos y, si los datos lo permiten, lesiones de jugadores.

  La implementación de estas pruebas puede ayudar a los entrenadores y equipos a optimizar sus estrategias antes y durante los partidos, maximizando las posibilidades de éxito, pero a la vez tenemos el desafío de contar la disponibilidad y calidad de los datos ya que es esencial para el éxito de las predicciones. La falta de datos históricos o datos incompletos puede afectar la precisión. La elección específica de las técnicas de machine learning se determinará más adelante en función de la viabilidad y eficacia.

Estas pruebas de conceptos se llevarán a cabo con el propósito de evaluar la viabilidad y el potencial impacto de estas tecnologías en el contexto del fútbol. Los resultados de estas pruebas proporcionarán información valiosa sobre cómo estas herramientas pueden ser aplicadas con éxito en la toma de decisiones y el análisis de rendimiento en el fútbol profesional.

4.5– Arquitectura general

En este apartado se presenta la arquitectura informática y metodología que se llevará a cabo para el desarrollo de las diferentes pruebas de conceptos.

4.5.1. Aspectos metodológicos

En esta sección, se detalla la metodología que guiará el desarrollo de este proyecto de análisis de datos y machine learning enfocado en el fútbol. Los aspectos clave de la metodología incluirán la selección de algoritmos, el diseño de experimentos, la evaluación de resultados. A continuación, se presenta un resumen de cómo se abordarán estos aspectos en el proyecto:

Selección de algoritmos y técnicas de machine learning

Se llevará a cabo una revisión minuciosa y exhaustiva de los diversos algoritmos de machine learning disponibles en la actualidad. El objetivo es identificar con precisión cuáles de estos algoritmos se ajustan de manera óptima a los objetivos particulares de este proyecto.

La selección de los algoritmos se fundamentará en su capacidad para abordar eficazmente los desafíos específicos planteados en el contexto del análisis y la predicción en el fútbol. Este proceso de selección asegurará que se utilicen las herramientas más adecuadas y efectivas para obtener resultados sólidos y relevantes en este emocionante campo de investigación.

Diseño del experimento

Se establecerá un enfoque claro para el diseño de experimentos. En el caso de predecir resultados de partidos, se definirán conjuntos de datos de entrenamiento y prueba que incluyan datos históricos de partidos anteriores. Además, se determinarán las métricas de evaluación que se utilizarán para medir el rendimiento de los modelos, como la precisión, el error cuadrático medio o la matriz de confusión. Esto garantizará un proceso de experimentación riguroso y bien estructurado.

Evaluación de resultados

Se establecerán procedimientos para la evaluación continua de los resultados. Esto implicará comparar las predicciones generadas por los modelos con los resultados reales de los partidos. Se considerará la utilización de técnicas como la validación cruzada para asegurar la robustez de los modelos. La evaluación se realizará de manera periódica y se documentarán los hallazgos.

Documentación

Se mantendrá una documentación completa y detallada de todas las metodologías utilizadas. Esto incluirá descripciones exhaustivas de las fuentes de datos, la preparación de datos, los modelos de machine learning y los resultados obtenidos. La documentación facilitará la replicabilidad y la comprensión del proyecto.

Tecnologías a utilizar: lenguajes y librerías Ide

Para llevar a cabo este proyecto de análisis de datos y machine learning con un enfoque en el fútbol, se ha llevado a cabo una minuciosa selección de las tecnologías, lenguajes y librerías más idóneas para satisfacer los objetivos y requisitos de la investigación. A continuación, se exponen en detalle las principales tecnologías que serán implementadas en este proyecto innovador:

Lenguaje de Programación Python

Python es ampliamente reconocido y utilizado en el campo de la ciencia de datos y el machine learning debido a su sintaxis clara y legible, lo que facilita el desarrollo y la depuración de código. Su versatilidad y la gran cantidad de recursos disponibles lo convierten en una elección sólida para proyectos en estos campos. Además, Python es respaldado por una comunidad activa de desarrolladores y científicos de datos, lo que significa que hay una amplia gama de soluciones y recursos disponibles en línea para abordar una variedad de desafíos en el proyecto.

Figura 4.1: Logo del lenguaje Python.

Jupyter Notebook con Anaconda

Se utilizará Jupyter Notebook, una plataforma interactiva de código abierto que facilita la creación y el intercambio de documentos que contienen código, visualizaciones y narrativas. Se instalará mediante Anaconda, que es una distribución popular de Python para la ciencia de datos que incluye muchas librerías y herramientas esenciales preinstaladas.

Figura 4.2: Logo de Jupyter y Anaconda.

Librerías Principales

Se aprovecharán diversas librerías específicas de Python para tareas de análisis de datos y machine learning, incluyendo, pero no limitándose a:

- **NumPy:** Para operaciones matriciales y numéricas eficientes.
- **Pandas:** Para la manipulación y análisis de datos estructurados.
- **Matplotlib y Seaborn:** Para la creación de visualizaciones y gráficos.
- **Scikit-Learn:** Para la implementación de algoritmos de machine learning.
- **TensorFlow o PyTorch:** Para tareas de deep learning, si es necesario.

Figura 4.3: Logo de Jupyter y Anaconda.

Control de Versiones (Git con GitHub)

La elección de estas tecnologías se basa en su idoneidad para tareas específicas, su popularidad en la comunidad de análisis de datos y machine learning, y su capacidad para satisfacer los objetivos del proyecto. La combinación de Python, Jupyter Notebook, Git/GitHub y las librerías mencionadas proporcionará un entorno robusto y flexible para el análisis de datos en el contexto del fútbol.

Figura 4.4: Logo de Git y GitHub.

4.6– Descripción de los datos utilizados

4.6.1. Origen y características de los datos

Primera prueba de concepto: Análisis de datos deportivos

- **Origen:** Los datos fueron obtenido desde StatsBombs[28], que es una empresa especializada en datos de futbol y análisis avanzado de futbol, los mismos han sido cargados en python mediante la librería statsbombpy, que permite utilizar los datasets públicos sin necesidad de tenerlos en físico.

- **Características:** El dataset es un dataset de eventos, donde se encuentran registrados todos los eventos del juego con todos los datos complementarios. A Continuación se detallan los atributos con los cuales se trabajó para implementar el análisis:

  - **Cantidad de registros:** 4750 registros.
  - **Números de columnas:** 87 columnas.

Detalle de los atributos más importantes para el análisis:

Atributo	**Descripción**
id	Identificador único para el evento.
player_id	Identificador del jugador involucrado en el evento.
team_id	Identificador del equipo al que pertenece el jugador.
duration	Duración o tiempo de duración del evento.
timestamp	Marca de tiempo que indica cuándo ocurrió el evento.
type	Tipo de evento, como gol, pase, disparo, etc.
shot_type	Tipo de disparo, por ejemplo, con el pie o de cabeza.
shot_outcome	Resultado del disparo, como gol, bloqueado, desviado, etc.
shot_technique	Técnica utilizada para el disparo, como volea, cabezazo, etc.
second	Segundo en el que ocurrió el evento.
minute	Minuto en el que ocurrió el evento.
pass_type	Tipo de pase, como pase corto, largo, centro, etc.
pass_outcome	Resultado del pase, como exitoso, interceptado, etc.
location	Ubicación en el campo donde ocurrió el evento.
pass_end_location	Ubicación en el campo donde finalizó el pase.
pass_body_part	Parte del cuerpo utilizada para el pase, como pie, cabeza, etc.

Tabla 4.1: Descripción de los atributos en el conjunto de datos, prueba 1.

Segunda prueba de concepto: Predicción de lesiones en atletas

- **Origen:** El conjunto de datos para la segunda prueba de concepto, fue obtenido de un repositorio público en la plataforma GitHub[30]. Lamentablemente, el conjunto de datos no proporciona información detallada sobre su fuente de origen, como el autor o la entidad responsable de recopilar los datos.

- **Características:** En este dataset se encuentran registradas todas las lesiones de los jugadores de una liga, los datos están distribuidos en tres archivos de datos, pero fueron unidos para simplemente tener uno.

  - **Cantidad de registros:** 2400 registros.
  - **Números de columnas:** 7 columnas.

Descripción de los conjuntos de datos:

- **Metric:** proporciona datos sobre la movilidad de la cadera y la compresión de la ingle de los 30 atletas. Registrando todos los días desde el 05/01/2016 al 30/04/2018.

- **GameWorkload:** Proporciona la carga de trabajo realizada por cada atleta en un día de juego.

- **Injuries:** Proporciona las fechas en las que los atletas se lesionaron durante un juego.

A continuación de desglosarán los atributos mas importantes y que fueron utilizados en el modelo predictivo:

Atributo	**Descripción**
Injury	Indica si un atleta sufrió una lesión durante un período específico (Sí/No).
Athlete_ID	Identificador único asignado a cada atleta en el conjunto de datos.
Date	Fecha en la que se registró la información relacionada con la lesión del atleta.
Game Workload	Medida de la cantidad de trabajo o esfuerzo físico experimentado por el atleta durante un juego.
Groin Squeeze	Evaluación de la movilidad o salud de la ingle del atleta (valor numérico).
Hip Mobility	Evaluación de la movilidad o salud de la cadera del atleta (valor numérico).
Rest Period	Cantidad de tiempo que el atleta tuvo para descansar y recuperarse después de un evento deportivo o una lesión (unidad de tiempo).

Tabla 4.2: Descripción de los atributos en el conjunto de datos, prueba 2.

Tercera prueba de concepto: Predicción de goles esperados

- **Origen:** Los datos fueron obtenido desde StatsBombs[28], que es una empresa especializada en datos de futbol y análisis avanzado de futbol, los mismos han sido cargados en python mediante la librería statsbombpy, que permite utilizar los datasets públicos sin necesidad de tenerlos en físico.
- **Característica:** El dataset es un dataset de eventos, donde se encuentran registrados todos los eventos del juego con todos los datos complementarios. Para esta prueba de concepto, hemos filtrado los campeonatos de LaLiga, en los periodos del 2012 al 2021, se tomaron los eventos de cada uno de los partidos de las competiciones y filtramos por los eventos de tiros al igual que las columnas relacionadas a tiros.
- **Cantidad de registro:** 16834 registros.
- **Números de columnas:** 15 columnas.

A continuación de desglosarán los atributos mas importantes y que fueron utilizados en el modelo predictivo:

Atributo	**Descripción**
shotAerialWon	Indica si el tiro fue ganado en el juego aéreo.
shotBodyPart	Describe la parte del cuerpo utilizada para el tiro.
shotFirstTime	Indica si el tiro se realizó de primera intención.
shotDeflected	Informa si el tiro fue desviado por un defensor.
shotOneOnOne	Indica si el tiro se realizó en una situación de uno contra uno con el portero.
shotOpenGoal	Describe si el tiro se realizó en una portería abierta.
shotOutcome	Muestra el resultado del tiro (gol, tiro al poste, etc.).
shotTechnique	Describe la técnica utilizada para el tiro.
shotType	Indica el tipo de tiro realizado (ejemplo: disparo desde dentro del área, disparo de falta, etc.).
playPattern	Describe el patrón de juego o la jugada que precedió al tiro.
x	Coordenada x en el campo donde se realizó el tiro.
y	Coordenada y en el campo donde se realizó el tiro.
goal	Variable binaria que indica si el tiro resultó en gol (1) o no (0).
Distance	La distancia desde donde se realizó el tiro a la portería.

Tabla 4.3: Descripción de Atributos.

4.6.2. Proceso de recolección y preparación de los datos

Primera prueba de concepto: Análisis de datos deportivos

En esta etapa se llevó a cabo la recolección y preparación de los datos, utilizando la biblioteca de StatsBomb. La biblioteca de StatsBomb proporcionó acceso a datos públicos, que se utilizaron como base para este análisis. En el proceso de preparación de datos, se aplicaron ajustes mínimos para garantizar que los datos fueran adecuados para su análisis.

En particular, se realizaron las siguientes acciones durante la preparación de datos:

- **Clasificación de pases completados**

  En el dataset se encontraron registros de pases completados con valores vacíos. Para asegurar la integridad de los datos y su utilidad en el análisis, se clasificaron estos pases como "completados".

```
#Cuando el campo de pass_outcome sea NaN lo sustituiremos por "Complete", a modo descriptivo
team1_team2['pass_outcome'].fillna('Complete', inplace=True)
team1_team2
```

Figura 4.5: Reemplazar valores del dataset.

- **División de coordenadas**

  Para facilitar el análisis y la visualización de los datos, se realizó una división de las coordenadas en dos columnas separadas. Esto resultó en una columna para el eje X (horizontal) y otra columna para el eje Y (vertical). Este enfoque permitió una representación más clara y detallada de las ubicaciones espaciales de los eventos en el campo.

- **División de coordenadas de fin de pase**

  De manera similar a la división de coordenadas de inicio de pase, se aplicó el mismo proceso a las coordenadas de fin de pase. Esta acción se realizó con el objetivo de poder generar gráficos precisos de las trayectorias de los pases.

```
pases[['x', 'y']] = pases['location'].apply(pd.Series)
pases[['finPase_x', 'finPase_y']] = pases['pass_end_location'].apply(pd.Series)
pases
```

Figura 4.6: División de coordenadas.

Segunda prueba de concepto: Predicción de lesiones en atletas

En cuanto a la recolección de datos, se obtuvieron tres archivos .csv de un repositorio en GitHub, que posteriormente se utilizaron como fuente de información. una vez descargado, el conjunto de datos fue utilizado para su posterior análisis y procesamiento. En particular, se realizaron las siguientes acciones durante la preparación de datos:

- **Unión de Datos**

  Unión de los tres conjuntos de datos en uno solo utilizando el identificador del Atleta como clave de unión. Esto permitió consolidar toda la información relevante en un solo conjunto de datos para un análisis más completo.

- **Creación de la columna de Lesiones**

  Creación de una columna adicional llamada "injuries" para especificar los días en los que hubo lesiones y los días en los que no las hubo. Esta información se obtuvo de un conjunto de datos separado que solo registraba los días con lesiones.

- **Separación de Métricas**

  División de una columna que contenía métricas combinadas, como "groin squeeze" y "hip mobility", en dos columnas separadas.

```
#Como tenemos una columna para medir dos metricas diferentes para los dos tipos metrica
new_metrics_df = metrics.pivot_table('value', ['athlete_id', 'date'], 'metric').reset_index()
```

```
new_metrics_df.head()
```

metric	athlete_id	date	groin_squeeze	hip_mobility
0	1	2016-05-01	297	36
1	1	2016-05-02	274	36
2	1	2016-05-03	291	56
3	1	2016-05-04	260	24
4	1	2016-05-05	284	35

Figura 4.7: Métricas.

- **Transformación de Datos de Fecha**

  Conversión del campo de fecha de tipo objeto a tipo fecha. Originalmente, este campo estaba en formato de objeto, lo que dificultaba el análisis temporal.

```
#Cambiamos a fecha la columna de fecha
injuries['date'] = injuries['date'].astype('datetime64[ns]')
injuries.dtypes
```

```
athlete_id            int64
date         datetime64[ns]
dtype: object
```

Figura 4.8: Transformación.

- **Cálculo de Días de Descanso**

  Introducción de una nueva columna que calculaba la cantidad de días de descanso que tenían los atletas antes de cada juego. Esto se basó en las fechas de los juegos y se utilizó para evaluar el impacto del descanso en el rendimiento.

```
#Agregamos la columna "rest_period" para calcular los dias de descanso
final_data['rest_period'] = final_data.groupby('athlete_id')['date'].diff()
```

```
final_data.head()
```

]:

	injury	athlete_id	date	game_workload	groin_squeeze	hip_mobility	rest_period
0	0	1	2016-05-05	402	284	35	4.0
1	0	1	2016-05-08	365	250	41	3.0
2	1	1	2016-05-11	457	331	33	3.0
3	1	1	2016-05-16	405	260	38	5.0
4	0	1	2016-05-20	407	378	60	4.0

Figura 4.9: Días de descanso.

- **Creación de Variables Dummy**

  Aplicación de variables dummy a la variable “Atleta Id”. Esto se hizo para eliminar dependencias ordinales y sesgos en el análisis. Las variables dummy se utilizaron para representar la información categórica de manera binaria.

```
#Creamos un dummy con la variable Athete_Id
dummy_variables = pd.get_dummies(final_data['athlete_id'])
ready_data = pd.concat([final_data,dummy_variables], axis=1)
```

Figura 4.10: Variable Dummy.

Tercera prueba de concepto: Predicción de goles esperados

En esta etapa se llevó a cabo la recolección y preparación de los datos, utilizando la biblioteca de StatsBomb. La biblioteca de StatsBomb proporcionó acceso a datos públicos, que se utilizaron como base para este análisis. En el proceso de preparación de datos, se aplicaron ajustes mínimos para garantizar que los datos fueran adecuados para su análisis. En particular, se realizaron las siguientes acciones durante la preparación de datos:

- **Transformación de la columna**

  Se identificaron las columnas que estaban vacías y se les asignó el valor "False" para garantizar que no se perdiera información crítica.

```
#Sustituimos los valores vacios por False
tiros_filt.shot_aerial_won = tiros_filt.shot_aerial_won.fillna(False)
tiros_filt.shot_first_time = tiros_filt.shot_first_time.fillna(False)
tiros_filt.shot_one_on_one = tiros_filt.shot_one_on_one.fillna(False)
tiros_filt.shot_open_goal = tiros_filt.shot_open_goal.fillna(False)
tiros_filt.shot_deflected = tiros_filt.shot_open_goal.fillna(False)
```

Figura 4.11: Transformación.

- **Aplicación de encode**

  Se abordó la tarea de codificar las columnas categóricas, incluyendo "shot_body_part", "shot_technique", "shot_type" y "play_pattern". Este proceso permitió convertir las categorías en representaciones numéricas, facilitando así su inclusión en modelos de aprendizaje automático.

```
#Hacemos Encoder para las variables categóricas

columns_to_encode = ['shot_body_part', 'shot_type', 'play_pattern']

# Crea una instancia de LabelEncoder
label_encoder = LabelEncoder()

# Aplica Label Encoding a las columnas seleccionadas
for col in columns_to_encode:
    tiros_filt[col] = label_encoder.fit_transform(tiros_filt[col])
```

Figura 4.12: Encoding de columnas.

- **Oversampling**

  Se realizó oversampling en la variable objetivo, ya que estaba desbalanceada. Para abordar este problema, se implementó una técnica de oversampling, que consiste en generar réplicas de las muestras de la clase minoritaria. Esto equilibró el conjunto de datos y mejoró la capacidad del modelo para aprender patrones de ambas clases.

```
#Hacemos el over sample para balancear la var
oversample_factor = len(df_majority) // len(df_minority)

# Realiza el oversampling
oversampled_minority = df_minority.sample(n=len(df_minority) * oversample_factor, replace=True, random_state=42)

# Combina el DataFrame oversampled_minority con la clase mayoritaria
df_oversampled = pd.concat([df_majority, oversampled_minority], axis=0)
```

Figura 4.13: Encoding de columnas.

- **Valores Atípicos**

  Fueron eliminados los datos atipicos de la columna "X"que eran mayores a 40.

```
#Eliminados estos valores atípicos

condicion = tiros_filt['Distance'] > 40

tiros_filt = tiros_filt[~condicion]
```

Figura 4.14: Eliminación de valores atípicos.

4.6.3. Metodología y enfoque de machine learning empleado

En estas pruebas de conceptos, se aplicará un enfoque de aprendizaje supervisado que se enfocará en la predicción de dos aspectos fundamentales en el mundo del fútbol: la determinación de si un tiro terminará en gol y la predicción de lesiones en los atletas.

Este enfoque se basa en el uso de algoritmos de clasificación, una rama del aprendizaje automático especialmente adecuada para abordar problemas de clasificación.

1. Selección de Algoritmos

La primera etapa de esta metodología implica la evaluación de diversos algoritmos de clasificación. Se analizarán en profundidad opciones como la regresión logística, K-Nearest Neighbors (KNN), árboles de decisión y modelos de redes neuronales.

Cada algoritmo será seleccionado y aplicado en función de su desempeño particular en el proceso de predicción. Esta fase es esencial para determinar cuál de estos algoritmos es el más apropiado para cada problema, teniendo en cuenta su capacidad para aprender patrones a partir de los datos de entrenamiento y hacer predicciones precisas.

Para la aplicación de los modelos de predicción de goles y predicción de lesiones, se optó por una estrategia de aprendizaje supervisado. Este enfoque se eligió debido a la disponibilidad de datos etiquetados que permitieron entrenar y evaluar los modelos en función de sus capacidades predictivas.

Específicamente, se emplearon varios algoritmos de aprendizaje supervisado para abordar estos problemas de predicción.

Los algoritmos seleccionados incluyen:

- **K-Vecinos Más Cercanos (KNN):** este algoritmo se basa en la idea de que las instancias similares tienden a tener etiquetas similares. Calcula la distancia entre los puntos de datos y clasifica un punto en función de la mayoría de las clases de sus k vecinos más cercanos.
- **Árboles de Decisiones:** son estructuras jerárquicas que se utilizan para tomar decisiones. En este contexto, se aplicaron para la clasificación y regresión, lo que los hace adecuados para ambos problemas.
- **Regresión Logística:** este algoritmo se utiliza comúnmente para problemas de clasificación binaria. Modela la probabilidad de que una instancia pertenezca a una de las dos clases y se ajusta mediante la optimización de parámetros.
- **Máquinas de Vectores de Soporte (SVM):** son excelentes para problemas de clasificación y regresión. Buscan encontrar un hiperplano que mejor separe las clases y pueden manejar datos no lineales mediante el uso de funciones kernel.

La elección de estos algoritmos se basó en su idoneidad para los problemas específicos, y cada uno de ellos fue evaluado en función de métricas de rendimiento como el accuracy y la F1-score, entre otras. Esta selección diversificada de algoritmos permitió explorar diferentes enfoques y determinar cuál de ellos funcionaba mejor para cada problema en particular.

2. Características Relevantes

En el proceso de análisis de datos, se llevará a cabo un minucioso estudio de las características que componen los conjuntos de datos. Este análisis tiene como objetivo identificar las variables más influyentes en la predicción de los resultados. Se utilizarán técnicas de selección de características que se basan en la importancia estadística y la correlación con el resultado del partido.

3. Entrenamiento y Evaluación

Una vez que se han seleccionado los algoritmos y se han definido las características más relevantes, se procederá al entrenamiento de los modelos. Este proceso implica el uso de un conjunto de datos históricos que contiene información detallada sobre partidos previos y sus resultados, así como datos sobre lesiones en atletas. Los modelos aprenderán de estos datos para realizar predicciones futuras.

Para evaluar la eficacia de los modelos, se utilizarán métricas de rendimiento como la precisión (accuracy) y el F1-score. La precisión mide la proporción de predicciones correctas realizadas por el modelo, mientras que el F1-score considera tanto la precisión

como la exhaustividad del modelo. Estas métricas proporcionarán una evaluación sólida del rendimiento de los modelos en términos de su capacidad para predecir con precisión el resultado de los partidos y la ocurrencia de lesiones en los atletas.

4.7– Diseño del experimento

Primera prueba de concepto: Análisis de datos deportivos

Objetivo: Desarrollar un análisis de datos deportivos, el cual genere información de los jugadores para la toma de decisiones.

1. Selección de Datos:

- Se utilizó un conjunto de datos que consta de 4750 registros y 87 columnas.

2. Selección de Atributos:

- Fueron filtrados los datos y tomadas columnas específicas de cuerdo al análisis a realizar.

3. Preprocesamiento de Datos:

Se realizaron siete cambios en la preparación de datos:

- Clasificación de pases.
- División de las coordenadas del evento.
- Creación del atributo fin pase.

Segunda prueba de concepto: Predicción de lesiones en atletas

Objetivo: Desarrollar un modelo predictivo para predecir la ocurrencia de lesiones en atletas.

1. Selección de Datos:

- Se utilizó un conjunto de datos que consta de 2,400 registros y 7 columnas.
- Se dividió el conjunto de datos en un 70 % para entrenamiento y un 30 % para prueba.

2. Selección de Atributos:

- Se utilizaron todos los atributos disponibles, ya que su número era manejable y no requería exclusión.

3. Preprocesamiento de Datos:

Se realizaron siete cambios en la preparación de datos:

- Unión de lo diferentes datasets.
- Creación de la columna de Lesiones.
- Separación de las métricas "groin_squeeze" y "hip_mobility".
- Transformación de Datos de Fecha.
- Cálculo de Días de Descanso de los atletas.
- Creación de Variables Dummy para el ID del jugador.
- Aplicación de oversampling para abordar el desequilibrio de clases.

4. Selección de Métricas de Evaluación:

Se utilizarán las siguientes métricas de evaluación:

- Exactitud (Accuracy): Para medir la proporción de predicciones correctas.
- Recall: Para evaluar la capacidad del modelo para identificar positivos verdaderos.
- F1-score: Para proporcionar una medida del equilibrio entre precisión y recall.

5. Selección de Algoritmos y Configuración de Hiperparámetros:

Se aplicaron los siguientes modelos de machine learning:

- K-Nearest Neighbors (KNN).
- Regresión Logística.
- Árboles de Decisiones.
- Soporte de Vectores (SVM).

6. Manejo de Desbalanceo:

Se implementó un proceso de oversampling para abordar el desequilibrio de clases en el conjunto de datos.

Este diseño del experimento garantiza que se apliquen técnicas adecuadas para desarrollar y evaluar modelos de predicción de lesiones en atletas. Las métricas de evaluación seleccionadas permitirán una evaluación completa del rendimiento de los modelos, y el oversampling ayudará a abordar el desequilibrio en la variable objetivo.

Tercera prueba de concepto: Predicción de goles esperados

Objetivo: Desarrollar un modelo predictivo para predecir la cantidad de goles en un partido de fútbol.

1. Selección de Datos:

- Se utilizó un conjunto de datos que consta de 16,834 registros y 15 columnas.
- Se dividió el conjunto de datos en un 80 % para entrenamiento y un 20 % para prueba.

2. Selección de Atributos:

- Se determinó la correlación de los atributos con la variable objetivo.
- Se eliminaron las columnas "y", "shot_technique" y "shot_first_time" debido a su falta de relación con la variable objetivo.

3. Preprocesamiento de Datos:

Se realizaron tres cambios de preparación de datos:

- Transformación de las columnas vacías a valores false.
- Aplicación de la codificación (encode) a las variables categóricas.
- Aplicación de oversampling a la variable objetivo para abordar el desequilibrio de clases.

4. Selección de Métricas de Evaluación:

Se utilizarán las siguientes métricas de evaluación:

- Exactitud (Accuracy): Para medir la proporción de predicciones correctas.
- Recall: Para evaluar la capacidad del modelo para identificar positivos verdaderos.
- F1-score: Para proporcionar una medida del equilibrio entre precisión y recall.

5. Selección de Algoritmos:

Se aplicaron los siguientes modelos de machine learning:

- K-Nearest Neighbors (KNN).
- Regresión Logística.
- Árboles de Decisiones.
- Soporte de Vectores (SVM).

6. Manejo de Desbalanceo:

Se implementó un proceso de oversampling para abordar el desequilibrio de clases, ya que había una gran diferencia en la cantidad de muestras para cada clase (14,000 con valor 0 y 2,000 con valor 1).

Este diseño del experimento asegura que los datos sean representativos y que se apliquen las técnicas adecuadas para desarrollar y evaluar modelos de predicción de goles en partidos de fútbol. Las métricas de evaluación seleccionadas permitirán una evaluación completa del rendimiento de los modelos.

4.8– Análisis de Resultados

Primera prueba de concepto: Análisis de datos deportivos

En este análisis de datos deportivos se logra informaciones muy valiosas para la toma de decisiones para las empresas y dueños de equipos sobre el desempeño de sus jugadores. Entre los resultados obtenidos están los siguientes:

1. Análisis de Posesión del Balón:

Durante esta prueba de concepto se ha analizado el porcentaje de posesión del balón de los equipos. Esta información es fundamental para entender cómo un equipo controla el juego en el campo. Los equipos que mantienen una alta posesión tienden a tener más oportunidades de ataque y pueden desgastar a sus oponentes.

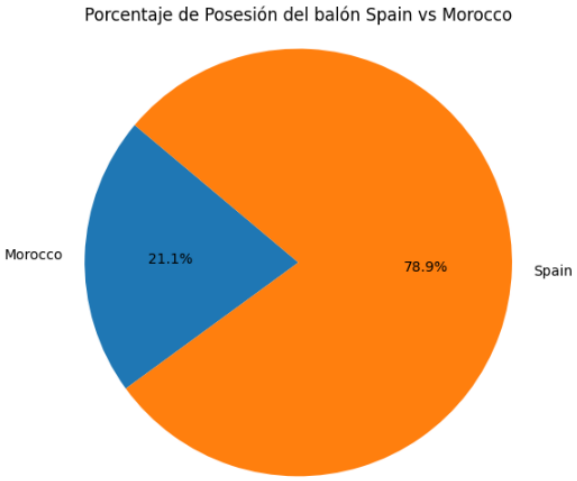


Figura 4.15: Gráfica del porcentaje de posesión del balón de ambos equipos.

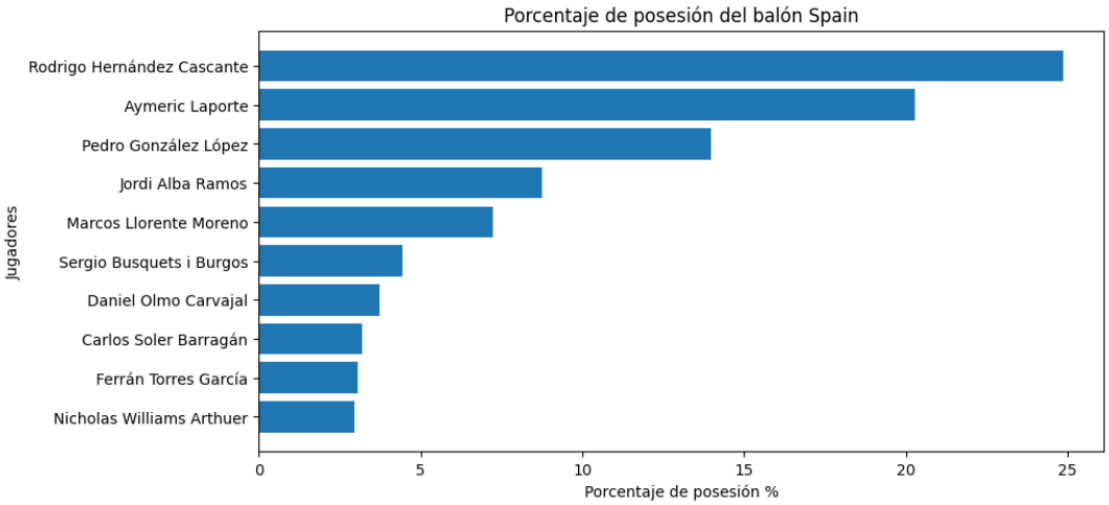


Figura 4.16: Gráfica del porcentaje de posesión del balón por jugadores.

2. Análisis de Efectividad de Pases de Jugadores:
Otro aspecto valioso de tu análisis fue la efectividad de los pases de los jugadores. Esto te permite identificar a los jugadores que son expertos en la distribución del balón y cuáles pueden necesitar mejorar en ese aspecto. Saber quiénes son los jugadores clave en términos de precisión en los pases es crucial para la estrategia del equipo.

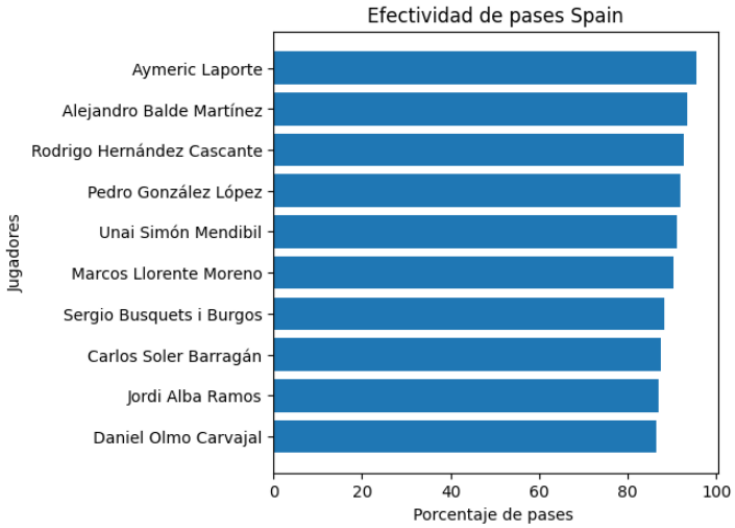


Figura 4.17: Gráfica de jugadores con mejor efectividad de pases.

3. Visualización de Zonas y Posiciones de Pases:
Las visualizaciones de las zonas y posiciones de los pases son extremadamente útiles para los entrenadores y analistas de fútbol. Estas gráficas te permiten identificar patrones de juego, como áreas del campo donde se concentran los pases exitosos o si los pases se originan desde una posición específica, como el mediocampo o las bandas. Estas visualizaciones pueden ayudar a ajustar tácticas y estrategias para aprovechar las fortalezas de tu equipo.

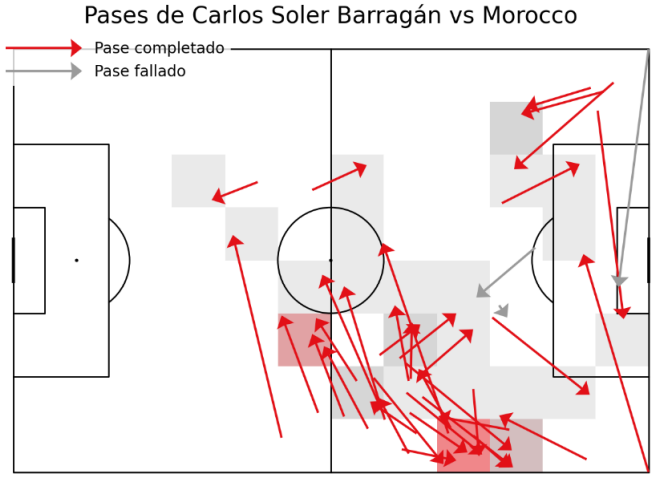


Figura 4.18: Representación gráfica de los pases completos e incompletos.

4. Completitud de Pases:

Además de la ubicación de los pases, la información sobre si los pases fueron completos o incompletos también es esencial. Esto puede revelar la consistencia y la precisión de los jugadores en la distribución del balón. Identificar a los jugadores que tienen un alto porcentaje de pases completos es fundamental para mantener la posesión y crear oportunidades de ataque.

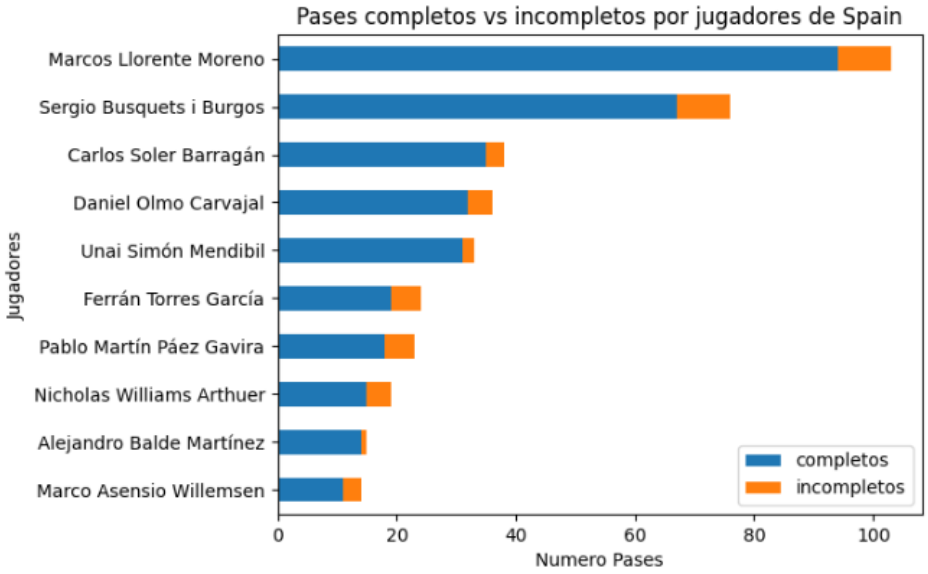


Figura 4.19: Gráfica de jugadores con mejor efectividad de pases completos vs incompletos.

En general, esta prueba de concepto proporciona información valiosa que puede ser utilizada por los equipos de fútbol para tomar decisiones estratégicas. Estos análisis pueden ayudar a los entrenadores a optimizar tácticas, identificar áreas de mejora y destacar a los jugadores clave en función de su efectividad en el campo. También demuestra la importancia del análisis de datos en el deporte moderno y cómo puede dar una ventaja competitiva a los equipos.

Segunda prueba de concepto: Predicción de goles esperados

En este análisis predictivo para la predicción de goles, se aplicaron cuatro modelos de clasificación: K-Nearest Neighbors (KNN), Regresión Logística, Árboles de Decisiones y Soporte de Vectores (SVM). Cada uno de estos modelos se evaluó utilizando métricas de precisión, recall y F1-score en dos clases (0 y 1), además de la precisión global (accuracy).

A continuación, se resumen y se comparan los resultados obtenidos de cada modelo:

1. K-Nearest Neighbors (KNN):

```
Accuracy: 0.93
[[2582  338]
 [  66 2843]]
              precision    recall  f1-score   support

           0       0.98      0.88      0.93      2920
           1       0.89      0.98      0.93      2909

    accuracy                           0.93      5829
   macro avg       0.93      0.93      0.93      5829
weighted avg       0.93      0.93      0.93      5829
```

Figura 4.20: Resultados KNN.

- Precisión global (accuracy): 93 %.
- Precisiones, recalls y F1-scores para ambas clases (0 y 1) son altos, indicando un buen equilibrio entre precisión y recall.
- Destaca por su alta precisión general y capacidad para predecir ambas clases de manera efectiva.

2. Regresión Logística:

```
Accuracy: 0.71
[[2003  917]
 [ 776 2133]]
              precision    recall  f1-score   support

           0       0.72      0.69      0.70      2920
           1       0.70      0.73      0.72      2909

    accuracy                           0.71      5829
   macro avg       0.71      0.71      0.71      5829
weighted avg       0.71      0.71      0.71      5829
```

Figura 4.21: Resultados Reresión Logística.

- Precisión global (accuracy): 71 %.
- Precisiones, recalls y F1-scores son moderados y similares para ambas clases.
- Obtiene una precisión global más baja en comparación con otros modelos, lo que sugiere un rendimiento equilibrado pero no excepcional.

3. Árboles de Decisiones:

```
Accuracy: 0.94
[[2598  322]
 [   4 2905]]
              precision    recall  f1-score   support

           0       1.00      0.89      0.94      2920
           1       0.90      1.00      0.95      2909

    accuracy                           0.94      5829
   macro avg       0.95      0.94      0.94      5829
weighted avg       0.95      0.94      0.94      5829
```

Figura 4.22: Resultados Árboles de Decisiones.

- Precisión global (accuracy): 94 %.
- Precisiones, recalls y F1-scores son altos para ambas clases.
- Destaca por su precisión excepcional y un buen equilibrio entre precisión y recall.

4. Soporte de Vectores (SVM):

```
Accuracy: 0.75
[[2049  871]
 [ 605 2304]]
              precision    recall  f1-score   support

           0       0.77      0.70      0.74      2920
           1       0.73      0.79      0.76      2909

    accuracy                           0.75      5829
   macro avg       0.75      0.75      0.75      5829
weighted avg       0.75      0.75      0.75      5829
```

Figura 4.23: Resultados SVM.

- Precisión global (accuracy): 75 %.
- Precisiones, recalls y F1-scores son moderados para ambas clases.
- Obtiene una precisión global aceptable pero inferior en comparación con los otros modelos.

En conclusión, en este análisis, el modelo de Árboles de Decisiones se destaca como el mejor clasificador para la tarea de predicción de goles. Con una precisión global del 94 % y altos valores de precisión, recall y F1-score para ambas clases, demuestra un rendimiento excepcional en la tarea.

K-Nearest Neighbors (KNN) también muestra un rendimiento sólido con una precisión global del 93 %, lo que lo convierte en una opción viable.

La Regresión Logística y el Soporte de Vectores (SVM) obtienen resultados menos impresionantes en comparación con los otros dos modelos.

Tercera prueba de concepto: Predicción de goles esperados

En este análisis predictivo para la predicción de lesiones, se aplicaron cuatro modelos de clasificación: K-Nearest Neighbors (KNN), Regresión Logística, Árboles de Decisiones y Soporte de Vectores (SVM). Cada uno de estos modelos se evaluó utilizando métricas de precisión, recall y F1-score en dos clases (0 y 1), además de la precisión global (accuracy).

A continuación, se resumen y se comparan los resultados obtenidos de cada modelo:

1. Regresión Logística:

```
Accuracy: 0.70
[[447 221]
 [177 492]]
              precision    recall  f1-score   support

           0       0.72      0.67      0.69       668
           1       0.69      0.74      0.71       669

    accuracy                           0.70      1337
   macro avg       0.70      0.70      0.70      1337
weighted avg       0.70      0.70      0.70      1337
```

Figura 4.24: Resultados Regresión Logística.

- Precisión global (accuracy): 70 %.
- Precisiones, recalls y F1-scores son moderados y similares para ambas clases.
- Obtiene un rendimiento equilibrado pero no excepcional.

2. Árboles de Decisiones:

```
Accuracy: 0.96
[[608  60]
 [  0 669]]
              precision    recall  f1-score   support

           0       1.00      0.91      0.95       668
           1       0.92      1.00      0.96       669

    accuracy                           0.96      1337
   macro avg       0.96      0.96      0.96      1337
weighted avg       0.96      0.96      0.96      1337
```

Figura 4.25: Resultados Árboles de Decisiones.

- Precisión global (accuracy): 96 %.
- Precisiones, recalls y F1-scores son muy altos para ambas clases.
- Destaca por su precisión excepcional y un buen equilibrio entre precisión y recall.

3. K-Nearest Neighbors (KNN):

```
Accuracy: 0.96
[[619  49]
 [  0 669]]
              precision    recall  f1-score   support

           0       1.00      0.93      0.96       668
           1       0.93      1.00      0.96       669

    accuracy                           0.96      1337
   macro avg       0.97      0.96      0.96      1337
weighted avg       0.97      0.96      0.96      1337
```

Figura 4.26: Resultados KNN.

- Precisión global (accuracy): 96 %.
- Precisiones, recalls y F1-scores son muy altos para ambas clases.
- Demuestra un rendimiento sólido y se encuentra en línea con Árboles de Decisiones.

4. Soporte de Vectores (SVM):

```
Accuracy: 0.99
[[658  10]
 [  0 669]]
              precision    recall  f1-score   support

           0       1.00      0.99      0.99       668
           1       0.99      1.00      0.99       669

    accuracy                           0.99      1337
   macro avg       0.99      0.99      0.99      1337
weighted avg       0.99      0.99      0.99      1337
```

Figura 4.27: Resultados SVM.

- Precisión global (accuracy): 99 %.
- Precisiones, recalls y F1-scores son muy altos para ambas clases.
- Obtiene un rendimiento excepcional con la precisión global más alta.

En conclusión, en la tarea de predicción de lesiones, los modelos de Árboles de Decisiones, K-Nearest Neighbors (KNN) y Soporte de Vectores (SVM) destacan por su excelente rendimiento. Todos estos modelos obtienen altos valores de precisión, recall y F1-score para ambas clases, lo que indica su capacidad para predecir con precisión tanto los casos de lesiones como los casos sin lesiones.

El modelo de Regresión Logística, aunque muestra un rendimiento equilibrado, se queda rezagado en comparación con los otros modelos.En términos de precisión global, el modelo de Soporte de Vectores (SVM) supera a los demás, con un impresionante 99 %.

4.9– Conclusiones sobre las pruebas de conceptos

En el proyecto realizado como resultado del trabajo de investigación, se ha llevado a cabo un estudio exhaustivo en el campo del análisis predictivo aplicado al fútbol. Se han explorado dos escenarios cruciales: la predicción de goles en partidos y la predicción de lesiones en jugadores. Durante este proceso, se aplicaron diversos modelos de machine learning con el objetivo de evaluar su capacidad para abordar estos problemas.

En términos generales, se pudo observar que la elección del modelo de machine learning es fundamental y debe adaptarse a las características específicas de cada problema. Algunos modelos sobresalieron en ciertos escenarios, mientras que otros demostraron ser más adecuados en diferentes contextos.

Este trabajo también pone de relieve la versatilidad y la aplicabilidad de las técnicas de machine learning en el ámbito deportivo, donde pueden desempeñar un papel crucial en la toma de decisiones, la estrategia y la optimización del rendimiento. Los resultados obtenidos en la predicción de goles y lesiones demuestran el potencial de estas herramientas para proporcionar información valiosa a entrenadores, equipos y profesionales de la salud en el deporte.

También, este trabajo representa un valioso aporte al campo del análisis predictivo en el fútbol y destaca la importancia de la selección adecuada de modelos y la evaluación rigurosa para abordar con éxito desafíos específicos en el mundo del deporte. Las lecciones aprendidas aquí tienen el potencial de beneficiar no solo al fútbol, sino a cualquier otro deporte o dominio que busque aprovechar el poder del machine learning para tomar decisiones basadas en datos y mejorar el rendimiento.

CAPÍTULO 5

Consideraciones finales

Con el presente trabajo de fin de master se ha llevado a cabo un estudio sobre el Machine Learning aplicado a los datos deportivos. Aquí se destacan las conclusiones principales que se han obtenido como resultado de este análisis, de igual forma se indican las contribuciones hechas, y las líneas de trabajo futuro que se pueden llevar a cabo.

5.1– Conclusiones

Este trabajo se fundamenta en los conceptos de la Inteligencia Artificial y el Machine Learning, para diseñar e implementar pruebas de conceptos que responda al dominio trabajado, que ha sido el de los datos deportivos en el Fútbol. De esta forma, se trata de facilitar ampliar y mejorar modelos estadísticos existentes sobre el mismo dominio.

Para un mayor control sobre el dominio tratado se han recopilado textos que sirven para cubrir la fase de adquisición del conocimiento, fundamental para la conceptualización de todo lo relacionado a sus reglas, formas de juego, y análisis de su situación actual.

Este trabajo presenta unas directrices claras para llevar a cabo el análisis de datos sobre cualquier dominio con las características descritas desde un punto de vista preliminar y en relación con una metodología concreta, fruto del análisis realizado.

Para llegar a la implementación de las pruebas de conceptos resultante se ha realizado un estudio sobre los principales elementos que toman acción durante un partido, encontrando así informaciones inherentes a los datos que se arrojan con cada acción de los jugadores, además de la evaluación del resultado final del juego.

La pruebas de conceptos realizadas se han llevado a cabo utilizando la herramienta Anaconda, con el lenguaje Python sobre Jupyter, ya que es ampliamente utilizado en distintos proyectos.

De forma particular:

- Se ha hecho un análisis previo de la actualidad del fútbol, así como del modelo predictivo a trabajar, definiendo aspectos como su dominio, datos a analizar y alcance.

- Se ha realizado una conceptualización del Machine Learning en el deporte, identificando los conceptos y elementos principales, así como las relaciones entre estos conceptos, sus características y sus atributos.

- Se ha formalizado dicha conceptualización, definiendo una prueba donde se aplican los elementos principales de IA y Machine Learning

Todo esto muestra la utilidad de la primera parte de este trabajo, que es la exploración documental del tema, para crear una base sólida sobre la que sustentar el análisis predictivo realizado. Adicionalmente, se puede identificar otro tipo de contribución en este trabajo, que es la implementación de las pruebas de conceptos, haciendo uso de las tecnologías más adecuadas para el análisis de los datos deportivos, que compone la segunda fase del proyecto realizado.

Lo más retador y a la vez interesante de realizar este proyecto ha sido primero conocer cómo funciona el mundo del fútobl, cuál es la realidad del Machine Learning en los partidos, y sobre todo, armar todas esas piezas para desarrollar las pruebas de conceptos que sustenta de forma práctica todo lo que se ha planteado de forma teórica en los primeros capítulos.

El futuro no está totalmente escrito, y en el deporte aún hay elementos que no se pueden medir, eso deja la puerta abierta a que el Machine Learning continúe brindando aportes significativos a todas las disciplinas, y en especial al Fútbol, deporte que invierte actualmente cantidades significativas para seguir mejorando la forma en que los partidos son analizados, y la manera en que son manejados los equipos.

Desde una perspectiva general, se considera que este trabajo aborda todos los objetivos planteados inicialmente, de igual forma cubre de forma estructurada la mayor parte de los elementos que se deben tomar en cuenta cuando se realiza un análisis de datos y cuando se desarrolla un modelo para predecir.

5.2– Contribuciones y limitaciones de la investigación

El desarrollo de este informe y sus pruebas de conceptos ofrecen una excelente fuente de conocimientos, tanto para entender la forma en que la IA ha ido cambiando la forma de juego, como para ofrecer medidas cuantitativas sobre la predicción de goles, de posibles lesiones de los jugadores, y del resultado que podría tener el partido, al entrenar un modelo con muchos de los datos que se generan durante un partido de fútbol.

Otro de los aportes resaltables del análisis realizado, es que aporta un modelo con muy buenos resultados en el campo de la predicción de diferentes acciones que ocurren durante el partido de fútbol.

Por otro lado, cabe resaltar que las prueba des conceptos solo se limita a dos aspectos fundamentales:

- Analizar los datos, cubriendo la cantidad de pases completados y no completados, la posición de los jugadores, posesión del balón, cálculo de la precisión en los pases.
- Crear pruebas de conceptos para predecir si un tiro se convertirá en gol, además de predecir lesiones en atletas.

5.3– Futuras líneas de investigación

Se sugiere la posibilidad de continuar futuras líneas de investigación que aborden el posible uso de Machine Learning para una mayor cobertura en el análisis de los jugadores, pero siempre considerando aspectos éticos, fundamentales para preservar la integridad de las personas.

El modelo predictivo realizado debería evaluarse y ponerse a prueba haciendo uso de otros datos, con los que se pueda evaluar su nivel de precisión y predicción sobre nuevos datos.

Para realizar un análisis exhaustivo sobre el desarrollo e implementación de otras pruebas de conceptos se considera interesante contar con el punto de vista de especialistas del mundo del fútbol, lo que permitirá refinar los puntos a cubrir en los modelos predictivos, además, contar con otros actores, como, entrenadores, jugadores y técnicos.

Finalmente, este trabajo se ha realizado teniendo en cuenta las limitaciones explicadas y cubriendo los objetivos planteados inicialmente.

Bibliografía

[1] Innovacion Gigital 360. Septiembre 2022. *Análisis de datos: Concepto, metodología y técnicas,* https://www.innovaciondigital360.com/big-data/analisis-de-datos-tecnicas-y-metodologias-para-la-aplicacion-de-analytics/.

[2] Alteryx. *Preparación de datos,* https://www.alteryx.com/es/glossary/data-preparation.

[3] IBM. *¿Qué es el análisis de datos exploratorio?,* https://www.ibm.com/es-es/topics/exploratory-data-analysis.

[4] Wikipedia, La enciclopedia libre. Julio 2023. *Selección de variable,* https://en.wikipedia.org/wiki/Feature_selection.

[5] Amazon. Abril 2015. *Amazon Machine Learning. Guía para desarrolladores,* https://docs.aws.amazon.com/es_es/machine-learning/latest/dg/training-ml-models.html.

[6] Sas. Abril 2021. *Inteligencia Artificial,* https://www.sas.com/es_cl/insights/analytics/what-is-artificial-intelligence.html.

[7] Fsm. Agosto 2022. *An Introduction to Machine Learning, Its Importance, Types, and Applications,* https://www.fsm.ac.in/blog/an-introduction-to-machine-learning-its-importance-types-and-applications.

[8] Tokio School. Abril 2023. *Conoce la historia del machine learning: ¡desde sus inicios!,* https://www.tokioschool.com/noticias/historia-machine-learning

[9] Salvador, M., Junio 2019. *Machine learning aplicado al Trading,* https://repositorio.comillas.edu/xmlui/bitstream/handle/11531/27863/TFG Salvador Maceira, Macarena.pdf?sequence=1isAllowed=y.

[10] Foqum. Julio 2023. *Introducción al Machine Learning,* https://foqum.io/blog/introduccion-al-machine-learning/.

[11] Oracle. Mayo 2021. *¿Qué es big data?,* https://www.oracle.com/es/big-data/what-is-big-data/.

[12] Alonso, C., *¿Qué es Open Data?,* Universidad de la Laguna, https://www.ull.es/catedras/catedrabob/que-es-open-data/

[13] Universidad de Puerto Rico. Marzo 2023. *Cómo encontrar datos sin procesar,* https://rcm-upr.libguides.com/c.php?g=1306512

[14] Power Data. *Data Warehouse: todo lo que necesitas saber sobre almacenamiento de datos,* https://www.powerdata.es/data-warehouse.

[15] Péres, I., Gegúndez, M., Abril 2021. *Deep learning : fundamentos, teoría y aplicación,* https://www.torrossa.com/it/resources/an/4947314.

[16] Westerbeek, H., Chmait, N., Diciembre 2021. *Artificial Intelligence and Machine Learning in Sport Research: An Introduction for Non-data Scientists,* https://www.frontiersin.org/articles/10.3389/fspor.2021.682287/full.

[17] Caressa, P., Noviembre 2021. *How Can AI support Football Tech Staff in Technical and Tactical Analysis and Decision Making?,* https://www.codemotion.com/magazine/ai-ml/ai-football-technical-tactical-analysis/.

[18] González, L., Diciembre 2020. *Inteligencia Artificial en los deportes,* https://aprendeia.com/inteligencia-artificial-en-los-deportes/.

[19] Issuu. Agosto 2023. *The Role of AI and Machine Learning in Sports Digital Marketing: Personalization and Beyond,* https://issuu.com/a2digi/docs/discover_how_ai_and_machine_learning_are_revolutio/s/29189312.

[20] Hansem, Jedd., Julio 2017. *Sports Performance Platform puts data into play – and action – for athletes and teams,* https://blogs.microsoft.com/blog/2017/06/27/sports-performance-platform-puts-data-play-action-athletes-teams

[21] Barcelona FC. Junio 2016. *Aplicando los principios de Johan Cruyff al data science,* https://www.fcbarcelona.es/es/ficha/1252109/aplicando-los-principios-de-johan-cruyff-al-data-science.

[22] Peranzo, P. Agosto 2023. *How Artificial Intelligence is Transforming the Sports Industry?,* https://imaginovation.net/blog/ai-in-sports-industry/.

[23] Shaker Mozn. Marzo 2023. *Cómo la Inteligencia Artificial está cambiando los deportes,* https://planetachatbot.com/como-inteligencia-artificial-esta-cambiando-deportes/

[24] LaLiga. Enero 2022. *LaLiga, pionera al estrenar en sus retransmisiones un modelo avanzado de Probabilidad de Gol graficado casi en tiempo real con tecnología de Microsof,* https://www.laliga.com/en-GB/news/laliga-takes-pioneering-step-by-adding-advanced-near-real-time-goal-probability-graphics-to-its-broadcasts-thanks-to-microsoft-technology

[25] Bartolomée, O., Noviembre 2021. *Creación de un Modelo de clasificación para calcular la métrica Xg,* https://www.maximaformacion.es/wp-content/uploads/2022/01/TFM_OscarBartolome.pdf

[26] Malagón, M., Septiembre 2019. *Machine Learning en el mundo del fútbol,* https://riunet.upv.es/bitstream/handle/10251/129491/Malagón - Machine Learning en el mundo del fútbol.pdf?sequence=1isAllowed=y

[27] Gómez, I., Junio 2021. *Estudio de técnicas de data science para la predicción de rendimientos deportivos,* https://repositorio.uam.es/bitstream/handle/10486/698384/gomez_carvajal_iñigo_tfg.pdf?sequence=1

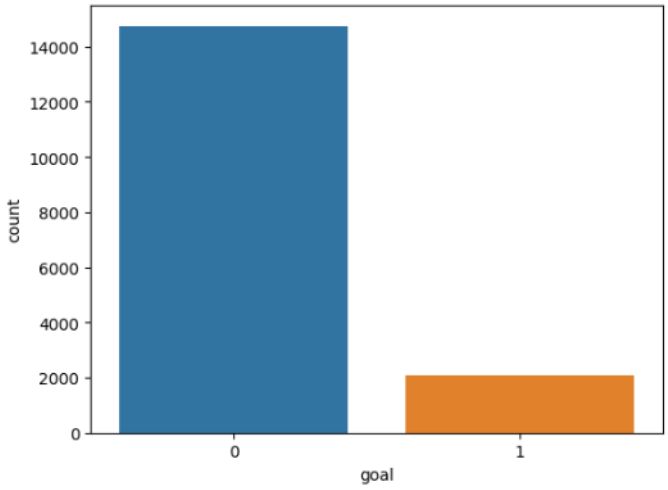
[28] Statsbomb. *Free Data,* https://statsbomb.com/what-we-do/hub/free-data/

[29] Kaggle. Junio 2021. *English Premier League stats 2019-2020,* https://www.kaggle.com/datasets/idoyo92/epl-stats-20192020

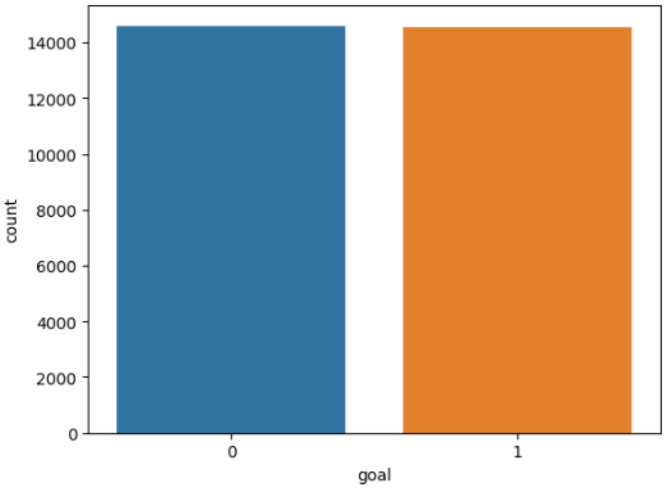
[30] Github. *Sports-Injury-Analysis* https://github.com/swathikiran86/Sports-Injury-Analysis

Anexos

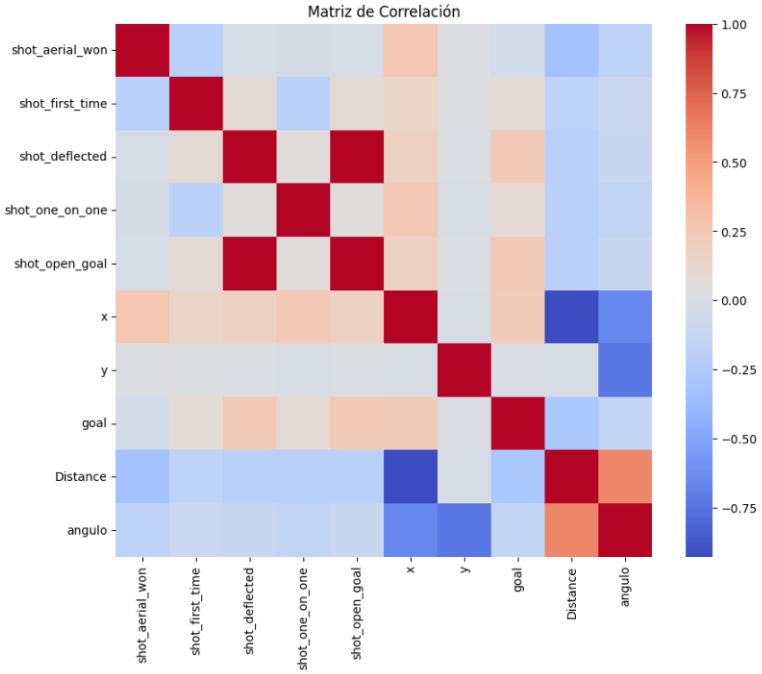
Anexo A. Distribución de la variable objetivo en la segunda prueba de concepto.



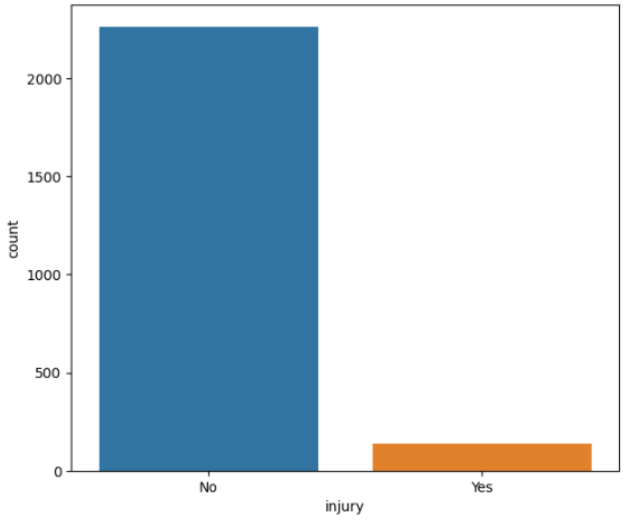
Anexo B. Distribución de la variable objetivo en la segunda prueba de concepto luego del oversampling.



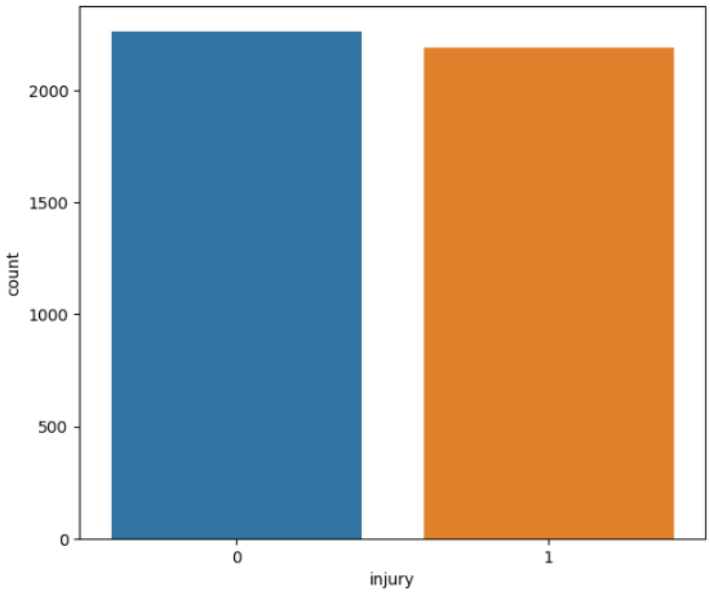
Anexo C. Matriz de correlación de la segunda prueba de concepto.



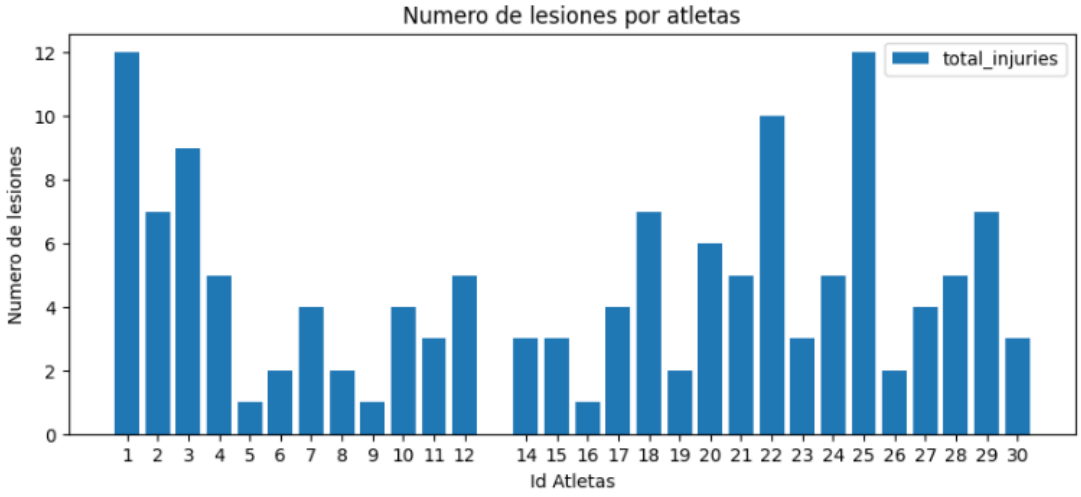
Anexo D. Distribución de la variable objetivo en la tercera prueba de concepto.



Anexo E. Distribución de la variable objetivo en la tercera prueba de concepto luego del oversampling.



Anexo F. Gráfica de la distribución de lesiones por jugadores.



Anexo G. Diagrama de Gantt del proyecto.