UNIVERSIDAD POLITÉCNICA SALESIANA

SEDE CUENCA

CARRERA DE TELECOMUNICACIONES

DESARROLLO DE UN DASHBOARD PARA MONITOREO DE VARIABLES FÍSICAS USANDO MÉTODOS DE IA PARA ANÁLISIS DE DATOS EN DEPORTES

Trabajo de titulación previo a la obtención del título de Ingeniero en Telecomunicaciones

AUTORES: FABIAN ANDRE MURILLO ALVARADO
JOEL FABRICIO QUINTUÑA VASQUEZ

TUTOR: ING. JUAN PAÚL INGA ORTEGA, MgT.

Cuenca – Ecuador

2025

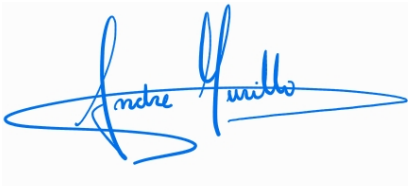
CERTIFICADO DE RESPONSABILIDAD Y AUTORÍA DEL TRABAJO DE TITULACIÓN

Nosotros, Fabian Andre Murillo Alvarado con documento de identificación N° 0605401025 y Joel Fabricio Quintuña Vasquez con documento de identificación N° 0104728886; manifestamos que:

Somos los autores y responsables del presente trabajo; y, autorizamos a que sin fines de lucro la Universidad Politécnica Salesiana pueda usar, difundir, reproducir o publicar de manera total o parcial el presente trabajo de titulación.

Cuenca, 10 de febrero de 2025

Atentamente,



Fabian Andre Murillo Alvarado

0605401025

Joel Fabricio Quintuña Vasquez

0104728886

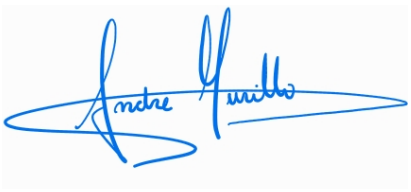
CERTIFICADO DE CESIÓN DE DERECHOS DE AUTOR DEL TRABAJO DE TITULACIÓN A LA UNIVERSIDAD POLITÉCNICA SALESIANA

Nosotros, Fabian Andre Murillo Alvarado con documento de identificación N° 0605401025 y Joel Fabricio Quintuña Vasquez con documento de identificación N° 0104728886, expresamos nuestra voluntad y por medio del presente documento cedemos a la Universidad Politécnica Salesiana la titularidad sobre los derechos patrimoniales en virtud de que somos autores del Proyecto técnico: "Desarrollo de un dashboard para monitoreo de variables físicas usando métodos de IA para análisis de datos en deportes" el cual ha sido desarrollado para optar por el título de: Ingeniero en Telecomunicaciones, en la Universidad Politécnica Salesiana, quedando la Universidad facultada para ejercer plenamente los derechos cedidos anteriormente.

En concordancia con lo manifestado, suscribimos este documento en el momento que hacemos la entrega del trabajo final en formato digital a la Biblioteca de la Universidad Politécnica Salesiana.

Cuenca, 10 de febrero de 2025

Atentamente,



Fabian Andre Murillo Alvarado

0605401025

Joel Fabricio Quintuña Vasquez

0104728886

CERTIFICADO DE DIRECCIÓN DEL TRABAJO DE TITULACIÓN

Yo, Juan Paúl Inga Ortega con documento de identificación N° 0104166491, docente de la Universidad Politécnica Salesiana, declaro que bajo mi tutoría fue desarrollado el trabajo de titulación: DESARROLLO DE UN DASHBOARD PARA MONITOREO DE VARIABLES FÍSICAS USANDO MÉTODOS DE IA PARA ANÁLISIS DE DATOS EN DEPORTES, realizado por Fabian Andre Murillo Alvarado con documento de identificación N° 0605401025 y Joel Fabricio Quintuña Vasquez con documento de identificación N° 0104728886, obteniendo como resultado final el trabajo de titulación bajo la opción Proyecto técnico que cumple con todos los requisitos determinados por la Universidad Politécnica Salesiana.

Cuenca, 10 de febrero de 2025

Atentamente,

Firmado electrónicamente por:
JUAN PAUL INGA ORTEGA

Ing. Juan Paúl Inga Ortega, MgT.
0104166491

AGRADECIMIENTOS

A mis padres, Elisa y Fabian por ser el pilar de todas las decisiones y valores; brindarme apoyo incondicional en cada decisión tomada; darme una niñez llena de mucho amor; y un millón de enseñanzas más que nunca terminaré de agradecer.

A mi hermana Maria Paz, por la cual siento una profunda admiración y respeto, quien me motivó día tras día a tener equilibrio en los pilares de vida; en quien veo el reflejo de la esencia que transmito; quien fue amiga y mentora en el desorden del día a día.

Al Ing. Juan Paúl Inga Ortega, MgT., mi tutor de tesis, por su guía experta, sus críticas constructivas, su disponibilidad incansable, visión y docencia salesiana. Y sobre todo por depositar su confianza en mi aún cuando no estuve seguro de mi mismo.

A mis amigos más cercanos Leslie, Jonathan, Anita, Belen, Domenica, Esthefy, Pablo, Ismael, Juan quienes han hecho ameno y lleno de alegría el transcurso de mi vida académica y personal.

A todas las personas que para bien y para mal fueron pasajeras en mi vida, porque sin ellas no hubiera nacido en mi la incertidumbre, la ansiedad o el miedo a lo desconocido, que fue combustible, incluso más que la propia motivación para superarme y aceptar el pasado.

Fabian Andre Murillo Alvarado

A mis padres, que siempre me han brindado su apoyo en mi formación académica. En especial a mi madre que siempre con su amor y paciencia ha estado en todos los momentos de mi vida apoyándome, sin el ejemplo trabajadora de ella este logro no hubiera sido posible.

A mis amigos, gracias por su compañía en todo este camino. En especial a mi mejor amigo Juan David, que estuvo apoyándome siempre y dándome consejos en todo este camino de la universidad, de igual manera a mi compañero de tesis Andre, le agradezco por su compromiso y dedicación para sacar este proyecto adelante.

A la personas que a lo largo de esta trayectoria pasaron por mi vida y que de alguna u otra manera me brindaron su apoyo para seguir adelante.

Agradecer a mi tutor de tesis Ing. Juan Paúl Inga Ortega, MgT. por su orientación, su conocimiento para guiarnos en el desarrollo de este proyecto y a lo largo de la carrera universitaria, nos permitieron llegar hasta este punto.

Por último y no menos importante, agradecerme a mi por creer en mi y nunca renunciar.

Joel Fabricio Quintuña Vasquez

DEDICATORIAS

Dedicatoria de Fabian Andre Murillo Alvarado

A mis padres, que me dieron la oportunidad de estudiar; a mi familia, que de forma constante me ha apoyado sin importar mis decisiones; a las personas que con gran cariño recuerdo y a quienes les hubiese gustado ver esta etapa de mi vida: Christian, abuelito Wilo; a mi yo del pasado, que vio con gran temor la etapa en la que me encuentro en este momento; a todas mis fieles mascotas, que cultivaron en mí la admiración por la naturaleza y el deleite de la tranquilidad, y pensé, serían eternas: Boffy, Timoteo, Abbie, Annie; a todas las personas que en su momento amé y me pidieron disfrutar del viaje sin apartar la vista del destino.

Joel Fabricio Quintuña Vasquez

A todos mis familiares y padres, que, sin importar las decisiones que tomé, me apoyaron y motivaron a concluir esta etapa de mi vida; a las personas que contribuyeron para que este proyecto se pudiera concluir; y a todas las personas que confiaron en mí, que supieron que algún día lo iba a lograr y a quienes, sin saberlo, también me impulsaron a llegar hasta aquí.

Índice general

Agradecimientos I

Dedicatorias III

Índice General V

Índice de figuras VII

Índice de tablas VIII

Resumen IX

Abstract X

Antecedentes 1

Justificación 4

Objetivos 5

Objetivo General 5

Objetivos Específicos 5

Introducción 6

1. Marco Teórico y Estado del Arte 8

1.1. Variables físicas y fisiológicas en el deporte 8

1.1.1. Deportes en equipo 9

1.1.2. Deportes individuales 9

1.1.3. Importancia del monitoreo de variables en el rendimiento deportivo 9
1.2. Conceptos básicos de IA y su aplicación en el deporte 10
1.2.1. Machine Learning 11
1.2.2. IA Generativa: Gemini 15
1.3. Estado del arte 16

2. Desarrollo, Integración del Modelo y Conexión a ThingsBoard 23
2.1. Modelos de ML Evaluados 25
2.1.1. Implementación de Modelos de ML para la Predicción de Lesiones 25
2.2. Envío y recepción de datos en tiempo real 36
2.3. Configuración y conexión con ThingsBoard 37
2.4. Implementación con la Api de Gemini. 39

3. Análisis de Resultados 42
3.1. Evaluación del Desempeño de los Modelos 42
3.2. Evaluación del Desempeño mediante Curvas ROC 43
3.3. Evaluación mediante Matrices de Confusión 45
3.4. Evaluación por métrica F1 47

4. Conclusiones, Recomendaciones y Trabajos Futuros 49

Glosario 53

Referencias 60

Índice de figuras

1.1. Análisis Bibliométrico: Mapa de Calor de Keywords en IA, ML y Deportes (1537 artículos). Fuente: Los Autores. 18
1.2. Red de Palabras Clave: Análisis Bibliométrico de IA, ML y Rendimiento Deportivo (1537 Artículos) Fuente: Los Autores. 18

2.1. Diagrama del proceso que realiza el proyecto. Fuente: El Autor. 24
2.2. Mapa de calor de la matriz de correlación. Fuente: Los Autores. 30
2.3. Correlación de todas las variables con el Riesgo de lesión.Fuente: Los Autores. 31
2.4. a) Distribución de la Edad. b) Probabilidad de Lesión por Grupo de Edad Fuente: Los Autores. 32
2.5. a) Distribución del IMC b) Relación con la Clasificación de IMC y Probabilidad de Lesión Fuente: Los Autores. 33
2.6. Distribución y análisis de la intensidad del entrenamiento en relación con la probabilidad de lesión. Fuente: Los Autores. 34
2.7. Distribución y análisis del tiempo de recuperación en relación con la probabilidad de lesión.Fuente: Los Autores. 35
2.8. Distribución de la variable *Likelihood_of_Injury*. Fuente: Los Autores. . 36
2.9. Dashboards de ThingsBoard. Fuente: Los Autores. 39

3.1. Curvas ROC obtenidas para los modelos evaluados. Fuente: Los Autores. 44
3.2. Matriz de confusión para distintos clasificadores. 46

Índice de tablas

1.1. Métodos supervisados. 12
1.2. Métodos no supervisados. 13
1.3. Métodos semi supervisados. 13
1.4. Métodos de Aprendizaje por Refuerzo. 14
1.5. Resumen de los trabajos relacionados con IA en el análisis de datos en el deporte. 22

2.1. Descripción de las variables del *dataset* 26
2.2. Estadísticas descriptivas de las variables analizadas. 26
2.3. Resumen de los tipos de datos, valores únicos y valores nulos. 27

3.1. Evaluación del rendimiento de los modelos de ML. 43
3.2. Valores de AUC obtenidos para cada modelo. 45

Resumen

Este trabajo identific al menos tres métodos para clasificar variables físicas y fisiológicas mediante aprendizaje automático (ML), con el objetivo de predecir posibles lesiones en deportistas. La selección del método de ML se basó en el análisis de las curvas ROC y las matrices de confusión. Los resultados obtenidos con la técnica de ML seleccionada alimentan una inteligencia artificial generativa (IA generativa) para proporcionar recomendaciones sobre la actividad física de los deportistas. Además, se desarrolló un dashboard en la plataforma ThingsBoard para monitorear variables físicas y fisiológicas y ofrece recomendaciones generadas por la IA. Los datos que a visualizar pueden ser recolectados por dispositivos IoT Entre las variables analizadas se incluyen la edad, el peso, la altura, lesiones previas, la intensidad del entrenamiento y el tiempo de recuperación. La metodología consistió en revisar la literatura sobre técnicas de IA en el deporte, seleccionar el mejor método para un objetivo común, implementarlo en Python con un dataset real, e integrar el modelo en un dashboard interactivo. Este dashboard sirve como herramienta de apoyo para entrenadores y especialistas en la toma de decisiones para la prevención de lesiones.

Palabras clave: Inteligencia Artificial; Machine Learning; dashboard; Monitoreo de variables físicas; deportes

Abstract

This work identifies at least three methods for classifying physical and physiological variables using machine learning (ML) to predict possible injuries in athletes. The ML method was selected based on the analysis of ROC curves and confusion matrices. The results obtained with the selected ML technique feed a generative artificial intelligence (generative AI) to provide physical activity recommendations for athletes. In addition, a dashboard was developed on the ThingsBoard platform to monitor physical and physiological variables and provide AI-generated recommendations. The data to be visualized can be collected by IoT devices. Variables analyzed included age, weight, height, previous injuries, training intensity, and recovery time. The methodology consisted of reviewing the literature on AI techniques in sports, selecting the best method for a common goal, implementing it in Python with a real dataset, and integrating the model into an interactive dashboard. This dashboard is a support tool for coaches and specialists in decision-making regarding injury prevention.

Keywords: Artificial Intelligence; Machine Learning; dashboard; monitoring of physical variables; sports.

Antecedentes

En *"Hábitos Atómicos"*, James Clear ilustra el concepto de las mejoras marginales mediante el ejemplo del equipo de ciclismo británico, que fue mediocre durante casi 100 años, en el año 2003, el entrenador Dave Brailsford implementó una estrategia de "mejora del 1 %", buscando pequeñas mejoras en cada área del equipo, desde la aerodinámica de las bicicletas hasta la higiene de los ciclistas, con el tiempo estos pequeños cambios acumulados llevaron a un éxito significativo, incluyendo la victoria de Bradley Wiggins en el Tour de Francia en 2012 y múltiples medallas olímpicas. Este ejemplo destaca que el éxito no siempre proviene de grandes saltos, sino de mejoras pequeñas y continuas que se acumulan con el tiempo [1].

De manera similar, el avance de la tecnología (IoT, del inglés *Internet of Things*) se ha basado en mejoras incrementales y continuas en diversas áreas tecnológicas. IoT, es un concepto que se refiere a la red colectiva de dispositivos físicos equipados con sensores y tecnología que facilita la comunicación entre dispositivos y nube para intercambiar datos a través de internet, como analogía, con el ejemplo del ciclismo, el éxito del IoT radica en la optimización progresiva de cada uno de sus componentes, desde los sensores inalámbricos hasta la nanotecnología. Esta mejora constante ha permitido la incorporación de *wearables* (tecnología portátil) y junto a ello su aplicación en una amplia gama de campos, como la salud, la energía, la biomedicina, la seguridad en la construcción, los deportes y la detección ambiental [2]-[6].

Los dispositivos *wearable* con sensores son una aplicación popular dentro de la IoT que ha atraído mucha atención en la última década, pueden monitorizar las actividades de un individuo durante el día y la noche, sin demasiadas interrupciones ni molestias [7]. En el ámbito deportivo, los sensores *wearable* han revolucionado

el monitoreo del rendimiento de los atletas, permitiendo la recolección y análisis de grandes cantidades de datos fisiológicos, mismos que una vez procesados y analizados, ofrecen información valiosa que puede ser utilizada para optimizar entrenamientos y prevenir lesiones [8].

En los últimos años la cantidad de datos ha aumentado de manera considerable como señala [9] en su artículo titulado "Tendencias de big data 2024: Navegando por el futuro de la tecnología de datos el volumen de datos"donde nos cuenta que la información recién generada, capturada, copiada o consumida, en los dos últimos años abarca el 90 % de los datos mundiales, esto ha llevado a la acumulación de grandes cantidades de información, el desglose de la creación de datos en distintos intervalos de tiempo expone el relieve de la magnitud del crecimiento que, gestionados de manera correcta, se traducen en información valiosa, sobre todo, en la industria del deporte donde han tenido un impacto sin precedentes. Los servicios de big data relacionados con el rendimiento, los datos fisiológicos, las estadísticas de entrenamiento y el análisis, han demostrado ser una ayuda eficaz en el entrenamiento diario y el desarrollo de estrategias de juego, y se están convirtiendo en un medio indispensable para ganar competiciones [10].

El análisis de datos subjetivos y objetivos puede personalizar el entrenamiento individual y prevenir lesiones, pero este proceso puede ser intensivo en tiempo y trabajo manual. Por esto es importante disponer de sistemas automatizados que puedan analizar grandes cantidades de datos procedentes de las sesiones de entrenamiento, y reflejen la información detallada a través de la visualización a favor de entrenadores y atletas.

Así, los cuadros de mando en inglés conocidos como *dashboards*, han emergido como herramientas para visualizar esta información, facilitando el proceso de monitoreo y la toma de decisiones [11], pues buscan centralizar y desplegar datos en forma que se puedan obtener ventajas operativas a través de una identificación rápida de las variables de interés, esto no solo reduciría los costos asociados al procesamiento manual de la información, sino que también permitiría agilizar la obtención de conclusiones y el diseño de estrategias de intervención o planes de entrenamiento más rigurosos encaminando a un mejor aprovechamiento de las variables físicas y

fisiológicas, pues como menciona [12] el uso de algoritmos inteligentes en sistemas centralizados reduce costos operativos y mejora la precisión en la toma de decisiones, decisiones basadas en evidencia que apoyen el rendimiento deportivo beneficiando a diversos actores, desde profesionales del deporte hasta investigadores y estudiantes.

Un ejemplo de esto es la Liga Americana de Baloncesto Profesional Masculino, donde se ha establecido un sistema completo de análisis de datos donde a través de big data, se rastrean las trayectorias de los jugadores, los árbitros y el balón, y se establecen indicadores de evaluación dinámicos que transforman estos datos en información estratégica para la mejora del rendimiento y planes estratégicos [13]. De forma similar, en Ecuador se han utilizado datos de baloncesto en base a fuentes ecuatorianas para analizar el desempeño de los jugadores mediante métodos estadísticos, lo que permite a los entrenadores obtener diversas interpretaciones que guían la toma de decisiones [14]. A nivel local se ha combinado hardware con tecnología IoT para capturar variables físicas y fisiológicas con monitoreo en tiempo real [15].

Justificación

Dada la creciente masificación de dispositivos *wearables* y sensores que miden indicadores como la frecuencia cardíaca, la intensidad del entrenamiento y el tiempo de recuperación, pero que carecen de una presentación unificada y eficiente de los datos recolectados, este trabajo propone el desarrollo de un *dashboard* para el monitoreo de variables físicas y fisiológicas de deportistas, con el objetivo de optimizar el análisis y la visualización de datos en tiempo real.

El *dashboard* desarrollado no solo facilita la visualización intuitiva de los datos para cada deportista registrado en una base de datos, sino que también incorpora métodos de IA para analizar la información y predecir riesgos de lesiones como factor adicional.

Así, a través del uso de IA generativa, el sistema sintetiza las respuestas de los clasificadores y las expone en el *dashboard*, permitiendo la generación de alertas tempranas y recomendaciones personalizadas; por supuesto quedando a discreción del experto que analice la información entregada para la toma de decisiones.

Además, la implementación de este sistema responde a la demanda de soluciones tecnológicas que integren grandes volúmenes de información de manera eficiente y ofrezcan análisis precisos en tiempo real. Además, el *dashboard* mejora la eficiencia en el monitoreo, beneficiando a investigadores, estudiantes y profesionales del deporte al proporcionar una herramienta que no solo simplifica el reconocimiento de tendencias, sino que también contribuye a una mejor comprensión del rendimiento y la prevención de lesiones.

Objetivos

Objetivo General

- Desarrollar un dashboard para monitoreo de variables físicas y fisiológicas usando métodos de IA para análisis de datos en deportes

Objetivos específicos:

- Establecer un estado del arte de métodos IA para el análisis de datos en deportes.
- Identificar los 3 métodos más adecuados conforme el estado del arte para su implementación.
- Construir un dashboard con el mejor método implementado.

Introducción

Según la Organización Mundial de la Salud (OMS) en [16], las lesiones deportivas son una causa relevante de discapacidad en atletas, mientras que el centro de innovación y conocimiento deportivo del FC Barcelona [17] reporta que ocurren 8,1 lesiones por cada 1.000 horas de exposición en deportistas de élite, impactando en su rendimiento y generando costos elevados para los clubes, esto ha reflejado un mayor interés por la medicina deportiva y el mercado global de wearables deportivos, valuado en $90.39 mil millones en 2024 [18], evidencia la demanda de soluciones tecnológicas que transformen datos en hallazgos que permiten tomar decisiones prácticas.

Se ha identificado que el impacto positivo del ejercicio físico en la salud está documentado desde un punto de vista fisiológico, mejorando la capacidad cardiovascular, regulando el peso corporal y contribuyendo al control de la presión arterial, también desempeña un papel crucial en la prevención y manejo de enfermedades crónicas como la diabetes tipo II, la obesidad y las enfermedades cardiovasculares, no solo a nivel físico sino más allá, en el ámbito de la salud mental, se ha demostrado que reduce los niveles de estrés, ansiedad y depresión, favoreciendo un mayor bienestar integral [19].

Con el avance tecnológico atletas y entrenadores han optado por escoger la visualización de datos digitales a través de monitores para manejar el desempeño, utilizando sistemas de seguimiento para dar soporte a su juicio de evaluación [20].

En la actualidad, las organizaciones deportivas enfrentan desafíos para integrar y visualizar estos datos, lo que limita la toma de decisiones preventivas. En este contexto, el presente trabajo aborda estas limitaciones combinando IA con plataformas como ThingsBoard y la API de Gemini, ofreciendo recomendaciones

personalizadas en tiempo real.

Es importante mencionar que, el estudio presenta restricciones: el dataset utilizado es limitado y no se consideran factores psicológicos o ambientales, a pesar de ello, la propuesta sienta las bases para optimizar el rendimiento deportivo, reducir lesiones y mejorar la gestión de datos, destacando la importancia de enfoques multidisciplinarios en futuras investigaciones.

Capítulo 1

Marco Teórico y Estado del Arte

Este capítulo establece los fundamentos teóricos y el contexto investigativo necesario para abordar el desarrollo del sistema propuesto, se inicia con un análisis de las variables físicas y fisiológicas en el deporte, diferenciando su relevancia en disciplinas individuales y de equipo, destacando la importancia de su monitoreo para optimizar el rendimiento y prevenir lesiones. Luego, se exploran los conceptos básicos de IA, enfocándose en técnicas como Machine Learning y la IA generativa, particularmente el modelo Gemini, y su aplicación en el ámbito deportivo. Al final se presenta un estado del arte que sintetiza investigaciones recientes mediante un análisis bibliométrico, identificando tendencias clave lo que permite contextualizar y justificar la selección de metodologías y herramientas utilizadas en el desarrollo del proyecto.

1.1. Variables físicas y fisiológicas en el deporte

Las variables físicas, relacionadas con la aceleración, la fuerza, la velocidad, la resistencia, agilidad, entre otras variables que reflejan las capacidades atlética; y variables fisiológicas, relacionadas con el estado de las funciones corporales, como frecuencia cardíaca, consumo de oxígeno, niveles de lactato, recuperación muscular, la temperatura corporal, etc. Son elementos fundamentales para comprender y optimizar el rendimiento deportivo, como explica Terrados en [21], la revisión de las bases fisiológicas comunes aporta información práctica para ajustar las cargas

de entrenamiento, conocer la situación metabólica de cada jugador durante la competición, y diseñar estrategias nutricionales y de recuperación de la fatiga.

1.1.1. Deportes en equipo

Los deportes de conjunto se caracterizan por ser acíclicos, con intervalos y discontinuos, lo que implica mantener la capacidad tanto aeróbica, como anaeróbica, durante la duración de las sesiones de juego. Aquí la variabilidad e imprevisibilidad del juego demandan una adaptación constante a situaciones tácticas y físicas, lo que exige la combinación de ejercicios de baja intensidad (como una carrera a baja velocidad) con actividades de alta intensidad (como *sprints* y saltos). Los parámetros antropométricos, junto con altos niveles de fuerza, potencia y velocidad, son factores clave para obtener una ventaja y lograr el éxito en los atletas de élite [22].

1.1.2. Deportes individuales

A diferencia de los deportes de conjunto, cuya naturaleza acíclica e impredecible demanda adaptación constante a estímulos externos, los deportes individuales se distinguen por patrones de movimiento especializados y repetitivos, donde el enfoque recae en la optimización de parámetros fisiológicos y biomecánicos predecibles. En disciplinas como el atletismo de velocidad o la natación, por ejemplo, el rendimiento depende de la capacidad para maximizar la eficiencia mecánica en movimientos cíclicos, como la fuerza aplicada en cada zancada o brazada [23], así como de la perfección técnica en gestos específicos [24]. Deportes como la gimnasia artística o el levantamiento de pesas enfatizan aún más la precisión técnica y el control neuromuscular, elementos que requieren un entrenamiento estructurado y menos condicionado por variables externas, como la interacción con adversarios [25].

1.1.3. Importancia del monitoreo de variables en el rendimiento deportivo

Es reconocido que el análisis avanzado del desempeño en deportes de alto rendimiento puede ser de ayuda para que el personal de apoyo, científicos del

deporte y profesionales aborden problemas complejos desde múltiples perspectivas [26], permitiendo el enfoque en equilibrar el rendimiento atlético óptimo con las cargas a las que están sometidos los deportistas, las cuales comprenden tanto las exigencias biológicas y psicológicas (carga interna o real) como las demandas físicas impuestas por las actividades deportivas (carga externa o propuesta), ya sea durante entrenamientos o competiciones. Con la incursión de las organizaciones deportivas en el ámbito de los macrodatos y gracias a la popularización de sistemas de tecnología portátil y sensores integrados en *wearables* que generan un flujo continuo de datos en tiempo real sobre el comportamiento, la adquisición y monitoreo resulta crucial para el tema de grandes volúmenes de datos (del inglés Big Data) deportivo, cuyo principal objetivo es el etiquetado y mejora de datos existentes limpiándolos y realizando tratamiento necesario según los distintos requisitos de la aplicación en cuestión para ofrecer heterogeneidad y estandarización en su recopilación [27], beneficioso para la accesibilidad centralizada y extensión de *datasets* de dominio público; también resulta de utilidad como alternativa al problema de privacidad y confidencialidad de los datos que se enfrenta mediante la generación de datos sintéticos por sus siglas SDG donde en esencia se estima un modelo estructural extraído de la "verdad fundamental" en los datos reales. Según la fuente de la que se parte [28], sugiere enfoques basados en el conocimiento donde la verdad se deriva de la teoría y expertos; basada en datos que se derivan de los datos reales a través de un mecanismo de estimación cuyo modelo generativo se construye con una amalgama de teoría, pero todas ellas utilizan como base un volumen grande que exhiban las mismas características estructurales de datos reales.

1.2. Conceptos básicos de IA y su aplicación en el deporte

La IA se basa en algoritmos inteligentes o de aprendizaje que permiten automatizar tareas y operaciones mediante algoritmos que, pueden entenderse como un conjunto preciso de instrucciones o reglas —o como una serie metódica de pasos— cuyo propósito es efectuar cálculos, resolver problemas y tomar decisiones de manera

eficiente. En la actualidad, la IA engloba tanto el aprendizaje de máquina (ML, del inglés *Machine Learning*) como el aprendizaje profundo (DL, del inglés Deep Learning) y se considera una pieza clave en la próxima revolución industrial, donde los datos han sido calificados como "el nuevo petróleo" [29]. Este fenómeno se observa en la rapidez con que se generan innovaciones tecnológicas, en el sector sanitario por ejemplo, la IA puede descubrir patrones antes desconocidos, acelerar procesos de diagnóstico y mejorar la calidad de los tratamientos. Aun así, su adopción en la medicina deportiva ha sido más lenta en comparación con ámbitos como las finanzas, la ciberseguridad, la manufactura o el comercio electrónico.

En el ámbito deportivo, se han adoptado de manera generalizada el uso de estadísticas y análisis de datos con el objetivo de medir y cuantificar diversos aspectos del comportamiento de los atletas, tanto dentro como fuera del terreno de juego. Siguiendo esta tendencia, la IA ha emergido como una herramienta clave para apoyar a los entrenadores en la toma de decisiones informadas y en la gestión de múltiples desafíos [26]. No obstante, la aplicación efectiva, la interpretación adecuada y el aprovechamiento de los datos en ciertos contextos relacionados con el comportamiento deportivo aún presentan limitaciones desafíos prácticos y éticos, como la privacidad, la seguridad y almacenamiento de datos, el intercambio y validez de los datos, y los problemas de replicabilidad [30]. Es importante señalar que más cantidad de datos no se traducen en una mejor calidad de la medición en la investigación.

1.2.1. Machine Learning

El ML es un conjunto de técnicas estadísticas diseñadas para desarrollar modelos predictivos precisos y replicables a partir de datos con múltiples variables dado que, en estos casos, los métodos tradicionales, como la regresión múltiple, pueden no ser efectivos para generar modelos confiables.

Para manejar grandes volúmenes de variables de manera eficiente, los algoritmos de ML exploran los elementos que pueden predecir o afectar los resultados y destacan aquellos que tienen mayor relevancia en la explicación de los resultados; aunque, estos métodos no garantizan la obtención de un modelo óptimo, sí permiten

identificar uno que funcione bien bajo diferentes condiciones. Debido a su naturaleza exploratoria, es fundamental aplicar técnicas para evitar el sobreajuste (*overfitting*), que ocurre cuando un modelo se ajusta demasiado a un conjunto de datos específico, dificultando su capacidad de generalización a nuevos datos. Además, grandes volúmenes de datos y variables posibilita la división de los datos en subconjuntos, permitiendo que algunos sean utilizados para el entrenamiento del modelo y otros para su validación, lo que ayuda a replicar los hallazgos dentro del mismo estudio. La precisión de los modelos automatizados también puede evaluarse comparándolos con análisis realizados por humanos en un subconjunto de datos [31]. En términos generales, los métodos de ML se pueden clasificar en cuatro categorías principales: aprendizaje supervisado, no supervisado, semisupervisado y por refuerzo. Cada una de estas categorías emplea enfoques y modelos distintos, cuyos detalles y definiciones se resumen en las tablas 1.1, 1.2, 1.3 y 1.4.

Métodos Supervisados

Funciona entrenando un modelo con datos etiquetados con el objetivo de encontrar las relaciones o estructuras en los datos de entrada que permitan al modelo generar las etiquetas de salida predefinidas [32]. Métodos como Máquinas de vector soporte (del inglés Support Vector Machines, SVM) y Redes Neuronales caen dentro de esta categoría y han sido empleados. Por ejemplo en [33] utilizaron Naïve Bayes, árboles de decisión y redes neuronales para la predicción del resultado de partidos y victorias en la NBA con una precisión del 83 %, analizando variables como porcentajes de tiros libres y rebotes.

Tabla 1.1: Métodos supervisados.

Objetivo	Métodos	Técnicas
Etiquetada (salida conocida) Tipo de variable: continua	Regresión	Regresión Lineal (Simple y Múltiple), Regresión Polinomial, Regresión LASSO y Ridge.
Etiquetada (salida conocida) Tipo de variable: categórica	Clasificación	Naive Bayes, Análisis discriminante lineal, Regresión Logística, KNN, SVM, Árbol de decisión, Random Forest, AdaBoost, XGBoost, SGD, Clasificación basada en reglas.
Etiquetada (salida conocida) Tipo de variable: numérica, dicotómica, categórica	Redes neuronales artificiales y DL	Perceptrón multicapa, Red neuronal convolucional, Memoria a largo plazo, Red neuronal recurrente.

Métodos No Supervisados

Aprende de los datos actuales no etiquetados sin intervención humana, y se utiliza para extraer características generativas, identificar tendencias, estructuras significativas, agrupar resultados y fines exploratorios [32]. Dentro de categorías como el clustering técnicas como K-means agrupan atletas según perfiles de rendimiento como en [34] donde se han utilizado en grupos o **clusters** para clasificar basquetbolistas identificando puntos fuertes y débiles que permitieron planes de entrenamiento individualizados.

Tabla 1.2: Métodos no supervisados.

Objetivo	Métodos	Técnicas
Sin etiquetar (salida desconocida)	Clustering	Métodos de particionamiento, Métodos basados en densidad, Métodos basados en jerarquía, Métodos basados en cuadrícula, Métodos basados en modelos, Métodos basados en restricciones, K-means, DBSCAN, Gaussian Mixture Models (GMMs).
	Aprendizaje de reglas de asociación	Sistema Inmunológico Artificial (AIS), Apriori, FP-Growth, ABC-RuleMiner.
	Reducción de dimensionalidad	Selección de características, Extracción de características, Umbral de varianza, Análisis de Componentes Principales (PCA), Eliminación Recursiva de Características (RFE), Selección basada en modelos.

Métodos Semi Supervisados

El aprendizaje semisupervisado combina una pequeña cantidad de datos etiquetados con una gran cantidad de datos no etiquetados [35].

Tabla 1.3: Métodos semi supervisados.

Objetivo	Métodos	Técnicas
Etiquetada de forma parcial (mezcla de salida conocida y desconocida)	Basados en grafos	Propagación de etiquetas, Propagación de creencias, Métodos de regularización basados en grafos.
	Métodos generativos	Modelos de mezcla gaussiana (GMMs), Expectation-Maximization (EM), Modelos generativos adversarios (GANs) semisupervisados.
	Basados en consistencia	Entrenamiento con consistencia de datos aumentados, Modelos basados en regularización de consistencia, Mean Teacher.

Métodos de Aprendizaje por Refuerzo

El modelo aprende a tomar decisiones mediante la interacción con un entorno, recibiendo recompensas o penalizaciones por sus acciones con el objetivo

de maximizar la recompensa a largo plazo y reducir el riesgo [36].

Tabla 1.4: Métodos de Aprendizaje por Refuerzo.

Objetivo	Métodos	Técnicas
Interacción con el entorno	Refuerzo positivo y Refuerzo negativo	Técnicas de Monte Carlo, Q-learning, R-learning.

Evaluación del desempeño

La evaluación común del rendimiento de un modelo se fundamenta en el uso de una matriz de confusión, lo que permite comparar la cantidad de predicciones acertadas y erróneas para cada categoría como se puede ver en las métricas de desempeño utilizadas en [33], [37]-[40], dentro de la matriz existen cuatro valores clave que deben ser analizados con atención:

- Verdadero positivo (TP): Corresponde al número de casos en los que el modelo predijo de forma correcta una observación como positiva [1].
- Falso positivo (FP): Representa la cantidad de veces que el modelo clasificó de forma incorrecta una observación negativa [2] como positiva
- Verdadero negativo (TN): Se refiere al número de casos en los que el modelo predijo de forma correcta una observación como negativa.
- Falso negativo (FN): Indica la cantidad de veces que el modelo clasificó de forma incorrecta una observación positiva como negativa.

Precisión global o Exactitud: En inglés *accuracy* se refiere a la medida de la proporción de predicciones correctas respecto al total de predicciones realizadas. Se expresa como un valor numérico que oscila entre 0 y 1, donde un valor de 1 representa un modelo con un rendimiento ideal, es decir, sin errores en sus predicciones.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{1.1}$$

[1]Positivo: se refiere a la clase de interés que el modelo intenta identificar (e.g., 'éxito', 'presencia' o 'clase A').

[2]Negativo: se refiere a la clase opuesta a la de interés (e.g., 'fracaso', 'ausencia' o 'clase B').

Precisión: En inglés *precision* evalúa la capacidad del modelo para clasificar de manera correcta las instancias pertenecientes a la clase positiva. En términos simples, esta métrica responde a la pregunta: de todas las predicciones que el modelo asigna a la clase positiva, ¿qué porcentaje de ellas son correctas? Si se utiliza la precisión como criterio para optimizar un modelo de forma única, se estaría priorizando la reducción de falsos positivos.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{1.2}$$

Exhaustividad o Sensibilidad: Del inglés *recall* evalúa la capacidad de un modelo para detectar de forma correcta todas las instancias positivas presentes en el conjunto de datos. Sin embargo, esta métrica no tiene en cuenta la cantidad de falsos positivos que el modelo pueda generar.

$$\text{Recall} = \frac{TP}{TP + FN} \tag{1.3}$$

Puntuación F1 Del inglés *F1 score* se define como la media armónica de precisión y recuperación.

$$F1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{1.4}$$

En su construcción, la puntuación F1 toma un valor entre cero y uno y podría expresarse como un porcentaje. Si la precisión y el recall son 1, la puntuación F1 también es 1, mientras que, la puntuación F1 es 0 cuando la precisión o el recall son 0.

1.2.2. IA Generativa: Gemini

Gemini es desarrollado por *Google DeepMind*, representa una de las arquitecturas más avanzadas de la IA generativa, se trata de un modelo de lenguaje grande (*Large Language Model*, LLM) diseñado para comprender y generar distintos datos a la vez, procesando información en distintos formatos como texto, imágenes y código, en su versión más avanzada, *Gemini Ultra 1.5*, introduce mejoras significativas en capacidad de razonamiento, eficiencia computacional y

precisión de respuestas a diferencia de otros modelos generativos que, dependen de los datos con los que fueron entrenados, *Gemini* se distingue por su capacidad de realizar consultas en tiempo real mediante *Google Search*, permitiéndo acceder a información actualizada y generar respuestas más precisas y contextualizadas. Incorpora mecanismos avanzados de evaluación de la calidad de la información, garantizando que las respuestas generadas sean verificables y confiables, desde el punto de vista arquitectónico, *Gemini* está basado en la familia de modelos *transformers*, optimizados para el procesamiento contextual y el aprendizaje profundo, lo que le permite comprender el contexto y aplicar razonamiento lógico en sus respuestas, convirtiéndolo en una herramienta versátil para múltiples aplicaciones, desde la asistencia en tareas empresariales hasta la generación automatizada de reportes técnicos [41].

1.3. Estado del arte

La calidad científica en el ámbito deportivo se ha convertido en un área de creciente interés, donde la investigación busca no solo mejorar el rendimiento de los atletas, sino que busca la integración de tecnologías avanzadas, en especial la IA, para optimizar su bienestar físico y mental, transformando la forma en que se entienden y gestionan los factores que afectan a los deportistas, desde el análisis de la carga interna de entrenamiento hasta la evaluación de factores externos propios del entorno, los estudios recientes destacan la interrelación entre aspectos físicos y fisiológicos.

Este estado del arte explora investigaciones recientes que emplean enfoques de IA para abordar desafíos comunes en diversas disciplinas deportivas, enfocándose en su aplicación para el rendimiento deportivo, para ello en esta sección se realiza una revisión literaria utilizando la base de datos digital *Web of Science (WoS)* donde se buscaron publicaciones y artículos relevantes anteriores al 2 de febrero de 2025 utilizando las palabras clave: (’machine learning’ OR ‘deep learning’ OR ‘predictive modelling’ OR ’artifcial intelligence’ OR ’ia’ OR ’data mining’ OR ’extreme learning machines’) AND (’team sport*’ OR ’sport’ OR ’performance’) donde los criterios de inclusión fueron: contener datos relevantes sobre IA en artículos que aborden el

uso de IA, ML o DL en deportes; que sean realizados en deportes individuales o en equipos; y su redacción se limite al idioma inglés. Los criterios de exclusión se limitan a no contener datos relevantes sobre IA aplicado a los deportes, cualquiera de las categorías: *early access*, *book chapters*, *proceeding papers* o *retracted publication*; y tener más de 3 años de publicación a la fecha como máximo.

La búsqueda inicial arrojó 1537 títulos de la base de datos elegida, los datos fueron importados al software de manejo bibliográfico *EndNote* donde 5 artículos duplicado fueron descartado dejando 1532, se filtraron por año dejando 808 artículos, los restantes se filtraron por palabras claves como: (sport OR sports OR 'Artificial Intelligence') dejando un total de 54 referencias, al final del procedimiento, se seleccionaron 11 artículos para su lectura en profundidad y análisis de métodos de IA utilizados, y se seleccionaron 10 de los más relevantes.

Se utilizó el software *VOSviewer* para construir y visualizar de forma global las redes bibliométricas de lo que la investigación nos dice sobre el uso de la IA en la comprensión del rendimiento deportivo, se agruparon los estudios según los temas de investigación más comunes en los deportes individuales y de equipo. Como los resultados mostraron en la figura 1.1, el conjunto de la investigación científica en este momento se ha centrado en: (1) predicción del rendimiento, (2) prevención de lesiones y (3) reconocimiento de patrones.

El mapa de redes basado en co-ocurrencias de palabras clave de la figura 1.2 indica un enfoque en el uso de IA con énfasis en aplicaciones para rehabilitación, clasificación, lesiones y desempeño utilizando variables externas como factores de riesgo, dando peso a los resultados de las investigaciones por métricas de fiabilidad de los modelos y concentrando el enfoque, en su mayoría en el fútbol.

En [42], se analiza la relación entre la carga interna de entrenamiento (TL), la recuperación y la disponibilidad en jugadores de fútbol profesional, utilizando enfoques de ML para comprender mejor cómo estas variables interactúan a lo largo de una temporada competitiva, en el estudio se subraya que la carga de entrenamiento afecta de forma negativa la recuperación y la disponibilidad de los jugadores, destacando la importancia de gestionar de manera adecuada las cargas para optimizar el rendimiento y reducir el riesgo de lesiones, menciona la aplicación

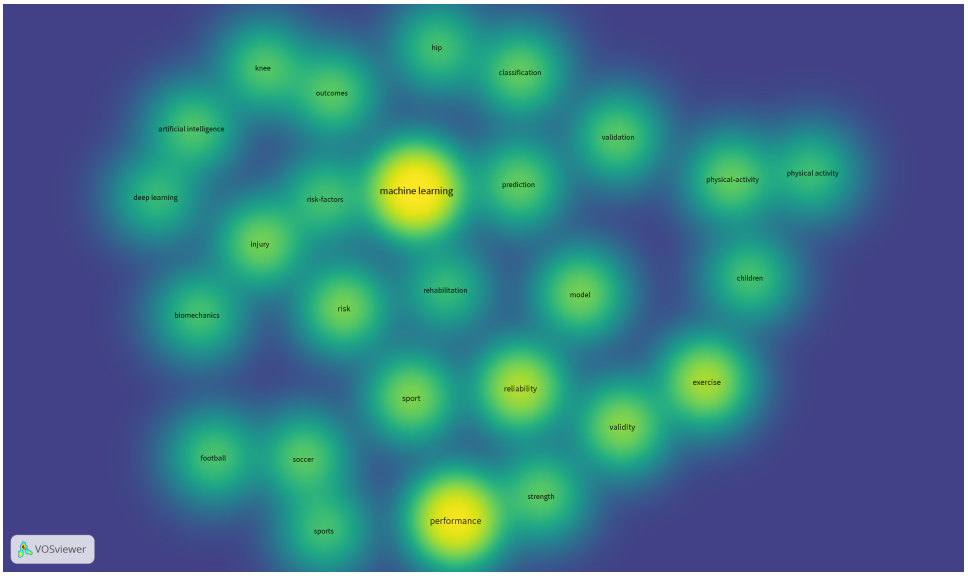


Figura 1.1: Análisis Bibliométrico: Mapa de Calor de Keywords en IA, ML y Deportes (1537 artículos).
Fuente: Los Autores.

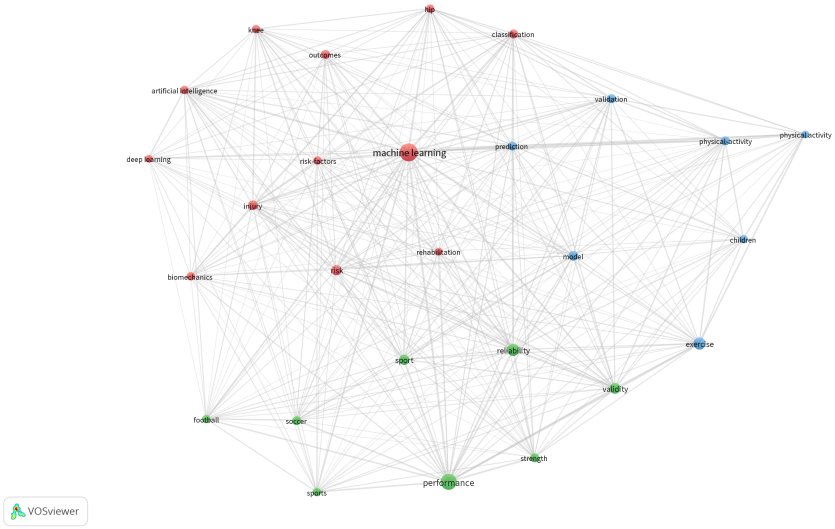


Figura 1.2: Red de Palabras Clave: Análisis Bibliométrico de IA, ML y Rendimiento Deportivo (1537 Artículos)
Fuente: Los Autores.

de técnicas como modelos de regresión para predecir la recuperación en función de la carga de trabajo, y análisis de series temporales para detectar patrones en los datos. Las métricas empleadas fueron el Índice de Esfuerzo Percibido o RPE para evaluar la carga interna y la escala TreS[3] para detectar información sobre la recuperación y

[3]TreS. Evaluación de Relaciones Transformacionales o Transformational Relationship Evaluation Scale

disponibilidad de entrenamiento; dichas métricas ofrecen una visión detallada del estado físico y fisiológico de los jugadores, resaltando la utilidad de herramientas de IA en la planificación y control de las cargas de entrenamiento en el deporte de élite.

Por otro lado, en [40], el artículo "Analysis and Prediction of Athlete's Anxiety State based on Artificial Intelligence" explora la comprensión y gestión del estrés psicológico en atletas, en particular en deportes de pista y campo, en este estudio se identifican las fuentes de presión psicológica y se desarrolla un modelo predictivo para analizar los estados de ansiedad mediante técnicas de IA utilizando un algoritmo de *clustering* jerárquico y una red neuronal de Funciones de Base Radial (del inglés Radial Basis Functions, RBF), los resultados obtenidos de 500 atletas demuestran que el modelo RBF supera en precisión a métodos tradicionales, lo que sugiere su potencial para mejorar las estrategias de manejo del estrés y, en consecuencia, el rendimiento deportivo. A pesar que, el enfoque principal es psicológico, también se subraya la importancia de integrar medidas fisiológicas en el reconocimiento de emociones y la evaluación del estrés, evidenciando la necesidad de un enfoque holístico que combine análisis psicológicos y fisiológicos.

En el ámbito del boxeo, [39] presenta un estudio que investiga el impacto de los estados psicológicos en la clasificación y reconocimiento de acciones de los boxeadores utilizando un modelo de IA. El estudio combina la recopilación de datos psicológicos a través de encuestas con un modelo de clasificación que fusiona un modelo DL de representaciones bidireccionales de codificadores a partir de transformadores por sus siglas en ingles BERT y una variante de las redes neuronales residuales por sus siglas en inglés 3D-ResNet, alcanzando una clasificación precisa de las acciones en el ring, con una precisión superior al 95 %, y se resalta la relevancia de los factores psicológicos en el rendimiento deportivo, proporcionando un marco robusto para el reconocimiento de acciones en el boxeo.

La investigación se extiende al análisis predictivo de lesiones de manera general, como se observa en [37], donde se plantea utilizar SVM utilizando el kernel de RBF con en técnicas de análisis de *big data* (Big Data Analytics, BDA), los resultados muestran que el modelo SVM propuesto logró una precisión del 92,3 % y una tasa de predicción del 87,5 % tras analizar tipos de lesiones, tiempos de recuperación,

tratamientos, frecuencia cardíaca, conteo de pasos, patrones de movimiento, minutos jugados, distancia cubierta y posición en el campo.

En deportes de carácter individual, [43] se enfoca en el tenis de mesa, utilizando técnicas de aprendizaje profundo, como redes neuronales convolucionales (del inglés Convolutional Neural Networks, CNN) y el algoritmo Adam, para reconocer movimientos técnicos con una precisión del 98.88 % analizando movimientos técnicos y la eficiencia del entrenamiento físico.

El trabajo de [44] presenta un modelo de ayuda en decisiones médicas para jugadores de fútbol basado en el análisis de datos históricos de lesiones mediante la técnica clasificador de árbol de decisión optimizado mediante selección de características y validación cruzada, como respuesta a la dificultad de aplicar algoritmos tradicionales de minería de datos en contextos médicos debido a la redundancia y desequilibrio en las categorías. Para ello, se midieron variables relacionadas con la carga de entrenamiento, incluyendo distancia total recorrida, aceleraciones, desaceleraciones, impacto de fuerza y carga metabólica arrojando resultados en los que el modelo alcanzó un 62 % de sensibilidad y un 42 % de precisión en la predicción de lesiones.

En este artículo [45] se analiza el uso de técnicas de Máquinas de Vectores de Soporte (Support Vector Classification, SVC) para la optimización de programas de entrenamiento de atletas seleccionando características mediante un algoritmo de computación evolutiva para selección de características relevantes mediante computación evolutiva optimizada (OA-EC-FS), las variables consideradas fueron: frecuencia cardíaca, intensidad del ejercicio, tiempo de recuperación, velocidad, fuerza, agilidad y niveles de estrés, evaluando la efectividad de los modelos propuestos mediante métricas de precisión, sensibilidad y puntuación F1.

El artículo [46] estudia el impacto del del término momento [4] en el rendimiento de los jugadores de tenis regresión logística múltiple, utilizando el algoritmo Light Gradient Boosting Machine (LGBM) para entrenar los datos y predecir el impulso en tiempo real de un atleta y máquinas de vectores de soporte (Support Vector Regression, SVR) para analizar los cambios de momento a lo largo del partido, lo

[4] Es un concepto tomado de la física (movimiento) y se aplica en el optimizador para actualizar los parámetros de los modelos de manera más eficiente.

que permitió identificar los factores más influyentes en el desarrollo del juego, donde se logró demostrar que el momento es real y afecta el resultado del partido, esto tras analizar las variables del entorno como puntuación, número de juegos ganados en un set, errores, velocidad de servicio, kilometraje, relación entre tiempo en la red y puntos ganados.

El artículo "A predictive analytics framework for forecasting soccer match outcomes using machine learning models" [38] desarrolla un análisis predictivo para predecir resultados de partidos de fútbol mediante modelos de ML. Se analiza el impacto de variables como resultados previos, estadísticas de juego, fatiga, momento y condiciones climáticas. Utilizando algoritmos como *Random Forest*, SVM, XGBoost, LightGBM y Redes Neuronales Convolucionales (CNN), combinados con técnicas de ensamblado como *Stacking* y *Voting*, donde se logró una precisión comparable a la de casas de apuestas, demostrando la utilidad de estos modelos en la predicción deportiva.

La literatura actual revisada en ML aplicada al fútbol se ha centrado en predecir resultados de partidos y detectar lesiones, dejando una brecha en el análisis del rendimiento individual basado en parámetros biométricos la cual es expuesta en [47] donde el problema está en la escasez de estudios que correlacionen datos biométricos y rendimiento individual, considerando las diferencias entre roles (delanteros, mediocampistas y defensores), y afrontadola mediante un enfoque novedoso que utiliza cuatro parámetros biométricos para predecir siete indicadores de rendimiento, identificando jugadores por encima del promedio del equipo. El logro principal es alcanzar una precisión superior al 90 % con Algoritmos como *Random Forest*, mejorada mediante una versión optimizada del Whale Optimization Algorithm (ED-WOA); Redes Neuronales Artificiales (ANN); Regresión Logística con AdaBoost (ADA-LR) y Árbol de Decisión con AdaBoost (ADA-DT). En cuanto a las variables medidas, comprenden parámetros biométricos (Costo energético, potencia metabólica, umbral anaeróbico, consumo máximo de oxígeno) e indicadores de rendimiento (distancia de aceleración, distancia de desaceleración, distancia recorrida) que permiten un análisis integral.

La convergencia de la IA y el análisis de datos en el deporte ofrece un marco

amplio, se puede ver un resumen de los trabajos revisados en la tabla 1.5.

Tabla 1.5: Resumen de los trabajos relacionados con IA en el análisis de datos en el deporte.

Artículo	**Variables Medidas**														**Objetivos**			**Métodos IA**	
Autor	Velocidad	Distancia Recorrida	Aceleración	Posición	Frecuencia Respiratoria	Frecuencia Cardíaca	Actividad Electrodérmica	Perfiles Físicos	Fuerza o Potencia Muscular	Percepciones Psicológicas	Perfil Metabólico	Tiempo de Recuperación	Agilidad	Variables del Entorno	Predicción del Rendimiento	Prevención de Lesiones	Reconocimiento de Patrones	ML: Aprendizaje Supervisado	ML: Aprendizaje No Supervisado
Pillitteri, et al., [42]	✠	✠	✠					✠	✠						✠			✠	
Guo., [40]					✠	✠	✠	✠	✠							✠		✠	✠
Kong y Duan., [39]										✠							✠	✠	
Li, W., [37]		✠		✠		✠						✠		✠			✠	✠	
Dongyang, [43]																	✠		✠
Fang, [44]	✠	✠	✠						✠		✠						✠	✠	
Zhang, [45]	✠				✠				✠	✠		✠	✠		✠			✠	
Tian, et al., [46]		✠												✠	✠			✠	
Wong, et al., [38]										✠				✠	✠			✠	
Morciano, et al., [47]					✠						✠				✠			✠	✠
Este trabajo	✠	✠						✠		✠					✠	✠		✠	

Capítulo 2

Desarrollo, Integración del Modelo y Conexión a ThingsBoard

Este capítulo discute la aplicación de técnicas de ML para predecir lesiones en deportistas mediante un conjunto de datos que contempla variables como la edad, el peso y la estatura de los jugadores; lesiones previas, intensidad del entrenamiento y tiempo de recuperación, con objetivo de desarrollar un modelo predictivo capaz de determinar la probabilidad de que un jugador sufra una lesión; lo cual resultaría valioso en términos de prevención y gestión de lesiones. En el transcurso del capítulo se describirá cómo se preparan los datos previo al análisis, se explicará la aplicación de distintos algoritmos de aprendizaje automático; también se examinará el rendimiento de los modelos resultantes.

En el desarrollo del dashboard para el monitoreo de variables físicas y fisiológicas en el ámbito deportivo se basó en la integración de herramientas tecnológicas, combinando técnicas de IA con una plataforma de gestión de datos en tiempo real. Este sistema permite analizar información clave de los jugadores y generar recomendaciones automatizadas para la prevención de lesiones.

La Figura 2.1 muestra el esquema que representa las etapas de adquisición, preprocesamiento, análisis y visualización de datos, asegurando una integración eficiente con los modelos de IA. A continuación se explica los pasos del esquema:

Se puede identificar que la primera fase implica la limpieza de datos que son obtenidos a través de un archivo con extensión **.CSV**. Antes de entrenar el

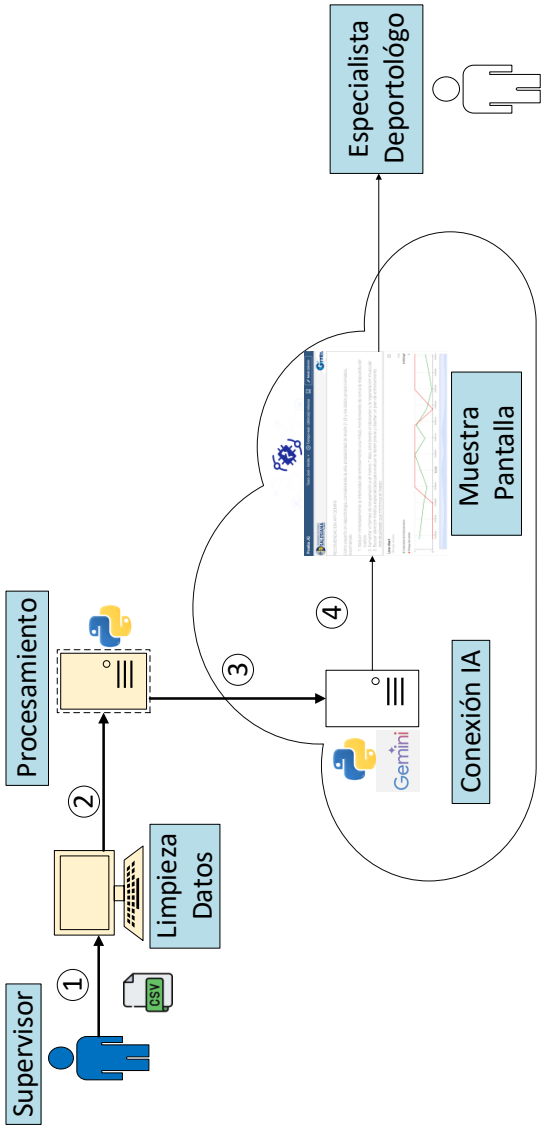


Figura 2.1: Diagrama del proceso que realiza el proyecto.
Fuente: El Autor.

modelo es necesario eliminar valores inconsistentes, datos incompletos y normalizar las variables asegurando compatibilidad con los modelos de ML.

En segunda instancia, se da la fase de procesamiento de datos. En este punto, los datos estos son cargados para su procesamiento en python, aplicando transformación, reducción de dimensionalidad y extracción de características con el objetivo de optimizar la eficiencia computacional. En esta etapa también, se implementan modelos de ML para determinar el mejor modelo a utilizar.

En la tercera etapa, se establece la conexión con la IA de Gemini a través de su

API para el análisis de datos de las predicciones. Esto se realiza através de un script hecho en Python. En el cual crea un prompt para las recomendaciones en tiempo real y se usa un temporizador para gestionar eficientemente el tráfico de datos y evitar inconvenientes con la API.

Los datos analizados son enviados a un dashboard en ThingsBoard, donde los resultados se presentan en formatos comprensibles, como gráficos y reportes, proporcionando una visualización clara para el usuario.

Las herramientas usadas para la realización del proyecto fueron, ThingsBoard que es una plataforma de IoT de código abierto utilizada para la recolección, visualización y gestión de datos. Python es el lenguaje de programación principal para el procesamiento de datos, conexión con APIs y envío de datos a ThingsBoard. Se seleccionó ya que contiene bibliotecas robustas y soporte para tareas de aprendizaje automático y análisis de datos. En cuanto a los frameworks web, aunque en esta etapa no se implementó directamente un *front-end*, ThingsBoard proporciona herramientas integradas para el diseño de paneles de control interactivos.

2.1. Modelos de ML Evaluados

En el proyecto, se evaluaron diferentes técnicas de ML para predecir lesiones en base al *dataset* recopilado, donde el proceso de selección del mejor modelo se realiza en base a métricas como *accuracy*, *recall* y *precision*, además de la exportación de los resultados en *excel* para su despliegue vía *dashboard*.

2.1.1. Implementación de Modelos de ML para la Predicción de Lesiones

El *dataset* utilizado en este proyecto proviene de la plataforma Kaggle[1] donde el conjunto de datos disponible en [48] recopila información relevante sobre atributos de jugadores y la probabilidad de sufrir lesiones. Cada registro en el *dataset* corresponde a un jugador e incluye las variables mostradas en la tabla 2.1.

[1]Kaggle es una plataforma de competencia de ciencia de datos y una comunidad en línea para científicos de datos y profesionales del aprendizaje automático de Google LLC.

Variable	Descripción
Edad del jugador	Edad del jugador en años.
Peso del jugador	Peso del jugador en kilogramos. Esta variable se distribuye con una media de 75 kg y una desviación estándar de 10 kg.
Altura del jugador	Altura del jugador en centímetros. Al igual que el peso, esta variable sigue una distribución normal con una media de 180 cm y una desviación estándar de 10 cm.
Lesiones previas	Indicador binario que señala si el jugador ha presentado lesiones previas (1) o no (0).
Intensidad de entrenamiento	Valor entre 0 y 1 que representa la intensidad de la sesión de entrenamiento del jugador.
Tiempo de recuperación	Número de días que se estima requiere el jugador para recuperarse de una lesión, con valores que oscilan entre 1 y 6 días.
Probabilidad de lesión	Indicador binario que refleja la probabilidad de que el jugador sufra una lesión (1) o no (0).

Tabla 2.1: Descripción de las variables del *dataset*

La estructura y el contenido de este *dataset* permite analizar de manera integral cómo las características individuales y los parámetros de entrenamiento pueden influir en la susceptibilidad a sufrir lesiones combinando variables cuantitativas para estudiar la relación entre características físicas, historial de lesiones y parámetros relacionados con la intensidad del entrenamiento.

Resumen estadístico

Las estadísticas resumidas del conjunto de datos pueden observarse en la tabla 2.2.

	Edad	Peso	Altura	Lesiones Previas	Intensidad Entrenamiento	Tiempo Recuperación	Probabilidad Lesión
Count	1000.0	1000.0	1000.0	1000.0	1000.0	1000.0	1000.0
mean	28.23	74.79	179.75	0.52	0.49	3.47	0.5
std	6.54	9.89	9.89	0.5	0.29	1.7	0.5
min	18.0	40.19	145.29	0.0	0.0	1.0	0.0
25 %	22.0	67.94	173.04	0.0	0.24	2.0	0.0
50 %	28.0	75.02	180.04	1.0	0.48	4.0	0.5
75 %	34.0	81.3	186.56	1.0	0.73	5.0	1.0
max	39.0	104.65	207.31	1.0	1.0	6.0	1.0

Tabla 2.2: Estadísticas descriptivas de las variables analizadas.

Se observa que las variables presentan una distribución variada en términos de escala y dispersión, la edad de los jugadores tiene un rango que va desde los 18 años hasta los 39 años, lo que indica una muestra que abarca tanto jugadores jóvenes como

veteranos. En cuanto al peso y la altura sugieren una diversidad en la composición física de los jugadores y variabilidad significativa en la estatura de los individuos analizados. En relación con las lesiones previas, el 51.5 % de los jugadores reportaron haber sufrido al menos una lesión, lo que subraya la importancia de considerar este factor en el análisis de riesgo. La intensidad del entrenamiento muestra una media de 0.49 indica que, en promedio, los jugadores realizan entrenamientos de intensidad moderada. Por otro lado, el tiempo de recuperación tiene una media de 3.47 días, con un mínimo de 1 día y un máximo de 6 días.

En términos de distribución, las variables analizadas presentan una dispersión moderada, con desviaciones estándar que oscilan entre 0.29 (para Intensidad) y 9.89 (para Peso y Altura), lo que sugiere que, aunque la mayoría de los datos se concentran alrededor de la media, existen valores atípicos identificados también como outliers [2].

Carga y Preprocesamiento de Datos

El *dataset* utilizado se carga desde un archivo CSV utilizando la función `pandas.read_csv()`. Antes de aplicar cualquier técnica de preprocesamiento, se realiza una exploración inicial de los datos, que incluye:

Inspección de los datos: Se examinan los tipos de datos, la cantidad de valores únicos y los valores nulos de cada columna lo que permite identificar posibles problemas como columnas con tipos de datos incorrectos o valores nulos como se muestra en la tabla 2.3.

	Tipo de dato	Únicos	Nulos
Edad	int64	22	0
Peso	float64	863	0
Altura	float64	875	0
Lesiones Previas	int64	2	0
Intensidad Entrenamiento	float64	101	0
Tiempo Recuperación	int64	6	0
Probabilidad Lesión	int64	2	0

Tabla 2.3: Resumen de los tipos de datos, valores únicos y valores nulos.

[2] outliers son datos anormales dentro de un conjunto de datos, es un valor extremadamente alto o bajo

Procesamiento de los datos: En el proceso de ingeniería de características dentro del análisis de datos, es común crear nuevas variables que capturen relaciones significativas entre los datos existentes, entre ellos el Índice de Masa Corporal (IMC) es una técnica reconocida en [49] que enriquece el conjunto de datos al transformar y combinar variables existentes en nuevas características más informativas y se calcula como se muestra en la ecuación 2.1.1:

$$\text{IMC} = \frac{W}{\left(\frac{H}{100}\right)^2}$$

Donde:

- W es el peso del jugador en kilogramos.
- H es la altura del jugador en centímetros.

Al clasificar el IMC en categorías como 'Peso Bajo', 'Normal', 'Sobrepeso', 'Obesidad I', 'Obesidad II' y 'Obesidad III', se facilita la identificación de patrones y tendencias que podrían no ser evidentes al considerar el peso y la altura por separado, esta categorización permite segmentar a los jugadores en grupos con características similares, lo que puede ser útil para análisis posteriores, como la evaluación del rendimiento deportivo en función de la composición corporal.

A continuación se emplea `OneHotEncoder` para transformar estas y más variables categóricas en representaciones numéricas, permitiendo su procesamiento en los modelos de ML. La función `train_test_split` divide el conjunto de datos en dos subconjuntos; el conjunto de entrenamiento, que se usa para entrenar el modelo de aprendizaje automático, y el conjunto de prueba, que se emplea para evaluar el rendimiento del modelo con datos que no ha visto antes, esencial para evitar el sobre ajuste (*overfitting*) y obtener una evaluación más precisa del modelo.

El coeficiente de correlación mide el grado en que dos variables cambian juntas y evalúa la relación monótona entre los distintos pares de variables. Se utilizó la correlación de Spearman para calcular la matriz de correlaciones y descubrir la correlación entre las distintas variables.

La figura 2.2 muestra la matriz de correlación identificada con colores semejante a un mapa de calor. La diagonal principal de la matriz presenta valores

de correlación iguales a 1, lo cual estaría correcto, dado que cada variable está correlacionada consigo misma. Se pueden identificar algunas relaciones importantes como por ejemplo, se observa una fuerte correlación positiva entre la variable *Player_Weight* y el Índice de Masa Corporal (IMC), con un coeficiente de 0.75, lo cual es consistente con la definición del IMC, que depende del peso del jugador. Además, existe una asociación significativa entre la variable *Clasificación_IMC* y *Player_Weight*, lo que indica que el peso del jugador influye en la categoría de su clasificación del IMC.

Hay variables con correlaciones bajas o cercanas a cero, lo que indica una relación lineal débil o inexistente, este análisis se lo realiza para la selección de variables en modelos de predicción, porque permite identificar relaciones que pueden influir en el desempeño del modelo.

Para ver la relación entre la probabilidad de lesión y el resto de las variables del conjunto de datos, se ha calculado la correlación de Pearson entre la variable *Likelihood_of_Injury* y las demás variables. La Figura 2.3 muestra los coeficientes de correlación de manera descendente, donde los valores positivos indican una relación directa con el riesgo de lesión, mientras que los valores negativos muestran una relación inversa.

El análisis indica que la variable con mayor correlación positiva con el riesgo de lesión es *Training_Intensity*, esto implica que una mayor intensidad de entrenamiento puede ser relacionada con un aumento en la probabilidad de sufrir lesiones. De manera similar, la variable *Grupo de edad_27-30* también muestra una correlación positiva, lo que indica que los jugadores en este rango de edad tienen un mayor riesgo de lesión en comparación con deportistas de otra edad.

Las variables como la clasificación *IMC_Obesidad II* presentan una correlación negativa con la probabilidad de lesión, lo que podría significar que jugadores con un índice de masa corporal más alto suelen tener tasas de lesiones bajas. Este comportamiento puede deberse a diversos factores fisiológicos o a diferentes estilos de juego y entrenamiento.

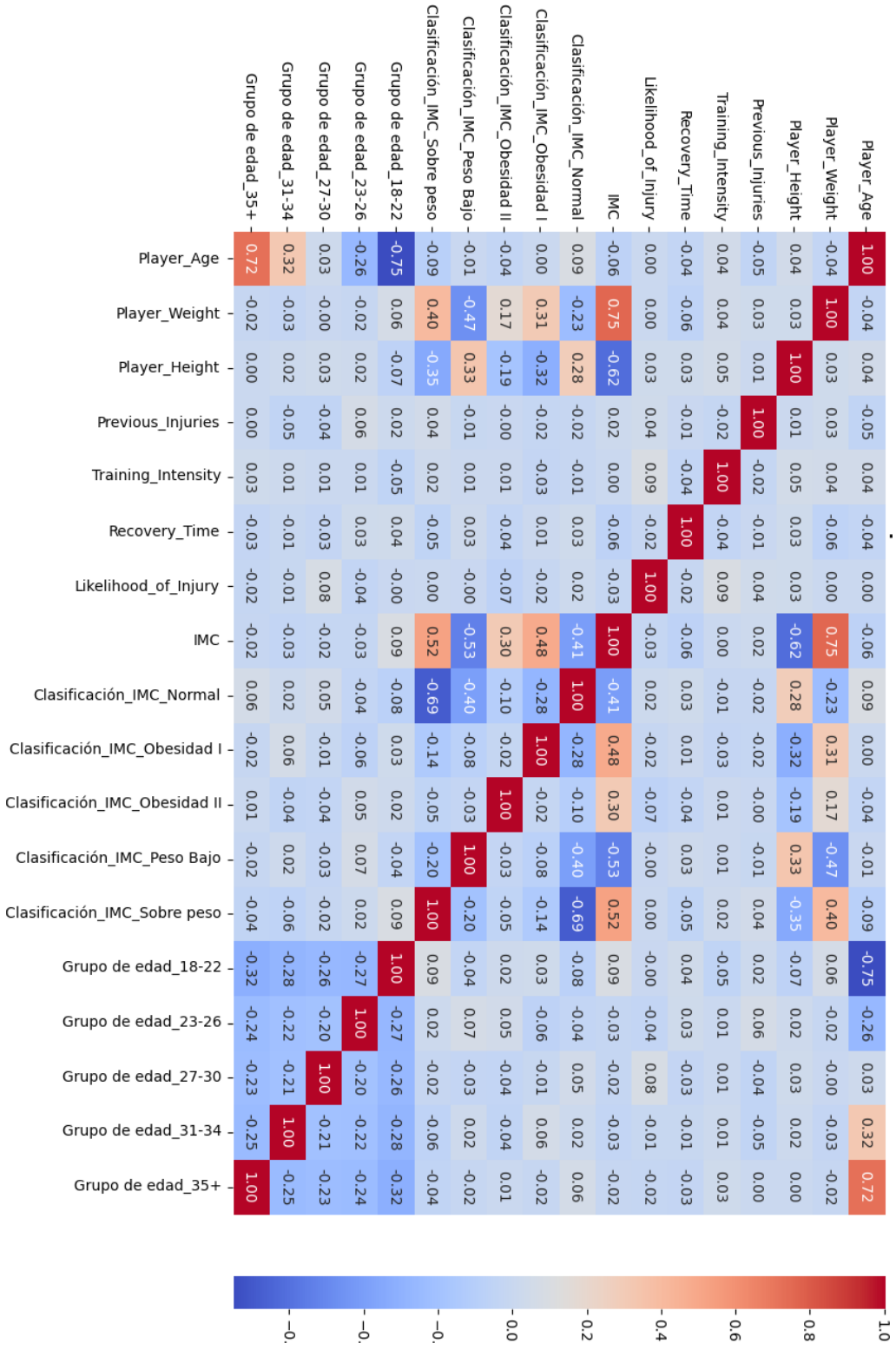


Figura 2.2: Mapa de calor de la matriz de correlación.
Fuente: Los Autores.

Normalización: Este es un proceso fundamental en el preprocesamiento de datos, cuando se utilizan algoritmos de ML que dependen de la escala de las características,

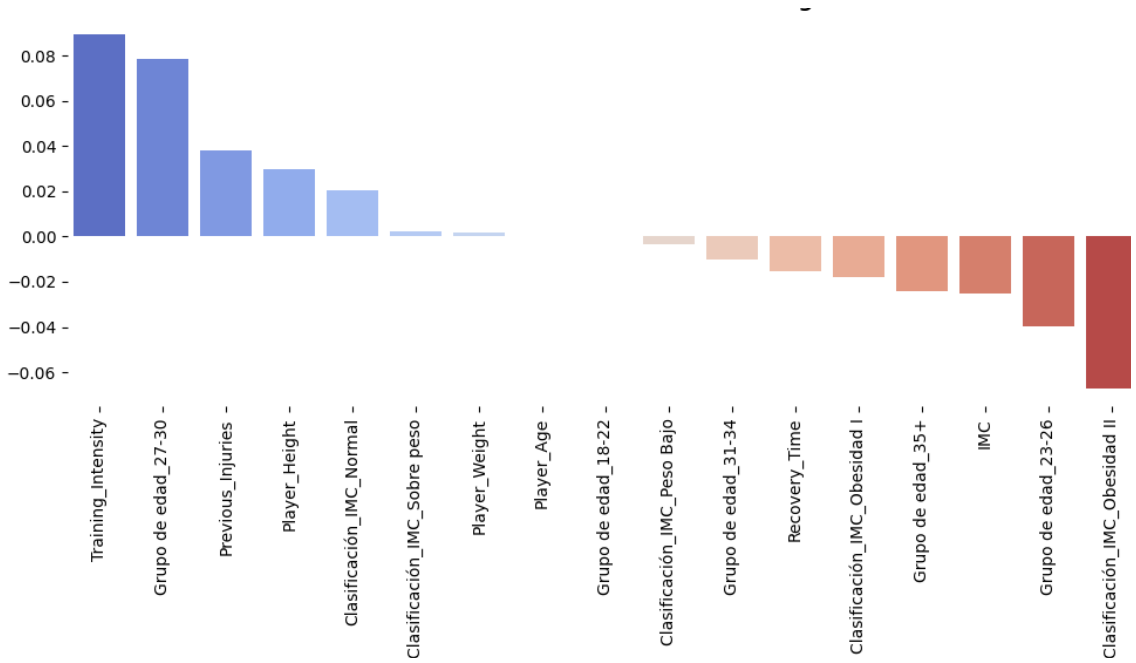


Figura 2.3: Correlación de todas las variables con el Riesgo de lesión.
Fuente: Los Autores.

como los basados en distancia. En este caso, se aplica el `MinMaxScaler` de `sklearn.preprocessing` para transformar las variables numéricas del conjunto de datos. El objetivo es reescalar las características a un rango específico, de forma típica [0, 1], preservando las relaciones entre las observaciones.

La normalización se realiza a través de los siguientes pasos:

1. Se importa `StandardScaler` y `MinMaxScaler` desde el módulo `sklearn.preprocessing`. En este caso, se utiliza `MinMaxScaler`, que transforma cada característica para que sus valores estén dentro del rango [0, 1].

2. Se crea una copia del dataframe original, `df_final`, bajo el nombre `df_final_normalizado`, con el fin de preservar el conjunto de datos original sin modificaciones.

3. Solo se seleccionan las columnas numéricas del dataframe, utilizando `select_dtypes()` para evitar problemas con variables categóricas que no pueden ser normalizadas.

4. Se aplica la normalización a las columnas numéricas seleccionadas mediante el método `fit_transform()`, el cual ajusta el escalador a los datos y los transforma

de acuerdo con el rango especificado (en este caso, [0, 1]).

Esto asegura que todas las variables numéricas estén en un rango similar, lo que mejora la convergencia de los algoritmos de ML y la comparación de diferentes características.

Comparación y exploración de variables

Para comprender mejor el conjunto de datos y la distribución de las variables, todas ellas se compararon con las observaciones de lesionados y no lesionados. Esta comparación permite reconocer las diferencias entre jugadores no lesionados y lesionados sin ningún algoritmo de aprendizaje automático.

- **Distribución de la edad y probabilidad de lesión por grupo de edad.**
  En este análisis se presentan los resultados de la distribución de la edad de los jugadores, y la relación entre la edad y la probabilidad de lesión. En la Figura 2.4, el primer gráfico, un *histograma*, muestra la distribución de las edades de los jugadores en el conjunto de datos

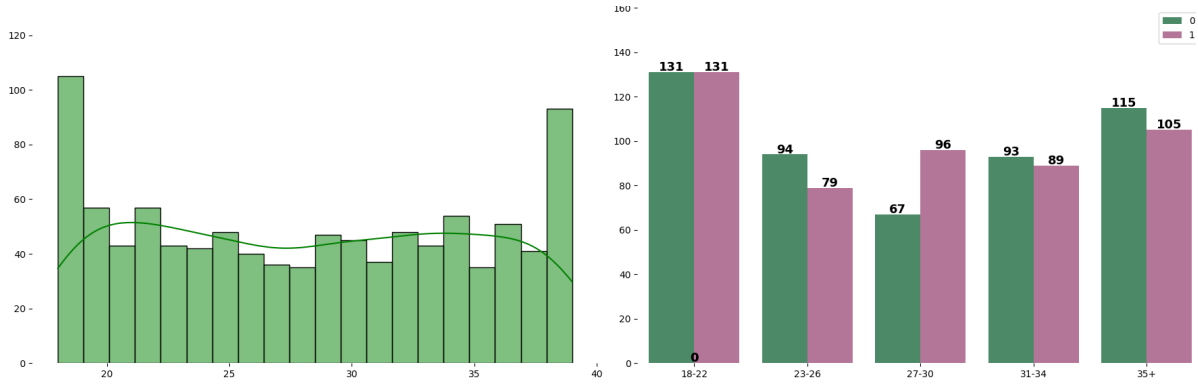


Figura 2.4: a) Distribución de la Edad. b) Probabilidad de Lesión por Grupo de Edad
Fuente: Los Autores.

A través de este análisis, es posible identificar el rango de edad más representado y obtener una visión general de cómo se distribuyen las edades dentro del conjunto de datos. El segundo gráfico, un *diagrama de barras*, compara la ocurrencia de lesiones entre los distintos grupos de edad. Agrupados en las categorías `18-22`, `23-26`, `27-30`, `31-34`, y `35+`. Cada barra representa un grupo de

edad y se utiliza un esquema de colores para diferenciar entre los jugadores con alta probabilidad de lesión (representados en rojo) y aquellos con baja probabilidad de lesión (representados en verde). Estos dos gráficos permiten explorar no solo la distribución de las edades, también la relación entre la edad y la probabilidad de lesión, lo cual ofrece información valiosa para estrategias de prevención de lesiones en el ámbito deportivo.

- **Análisis del Índice de Masa Corporal (IMC) y su relación con la probabilidad de lesión.**

  La gráfica mostrada en la Figura 2.5, presenta dos análisis importantes sobre el Índice de Masa Corporal (IMC) y su relación con la probabilidad de lesión. En la primera gráfica, se observa el histograma de la variable IMC, el cual permite visualizar su distribución a lo largo de los datos. Se puede notar que los valores del IMC están concentrados en un rango específico, con una mayor densidad en los valores bajos y medios, lo que sugiere que la mayoría de los jugadores tienen un IMC dentro de un rango considerado saludable. En el segundo gráfico, se presenta un gráfico de barras que compara la clasificación del IMC con la probabilidad de lesión. Cada barra representa un grupo dentro de la clasificación del IMC (Bajo Peso, Normal, Sobrepeso, Obesidad) y la probabilidad de lesión (0: baja probabilidad, 1: alta probabilidad). Este análisis muestra cómo la clasificación del IMC podría estar asociada con un mayor riesgo de lesión.

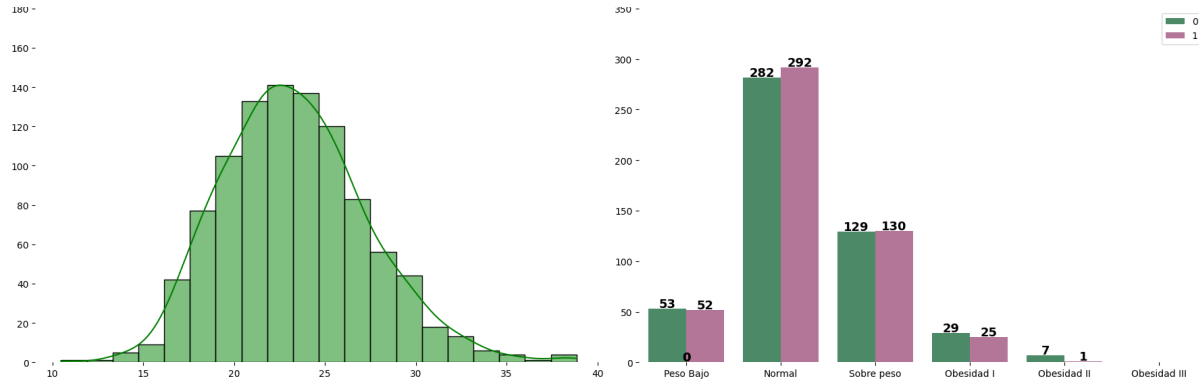


Figura 2.5: a) Distribución del IMC b) Relación con la Clasificación de IMC y Probabilidad de Lesión
Fuente: Los Autores.

- **Análisis de la intensidad de entrenamiento y su relación con la probabilidad de lesión.**

  En la gráfica 2.6 el histograma permite visualizar la distribución de la intensidad del entrenamiento en los deportistas analizados, proporcionando información sobre los distintos niveles de intensidad. Adicional, el gráfico de densidad que utiliza la variable *Likelihood_of_Injury* facilita la identificación de posibles diferencias en los patrones de entrenamiento entre jugadores con y sin historial de lesión. Por último, el diagrama de caja y bigotes ilustra la dispersión de los valores de *Training_Intensity*, destacando la mediana, el rango intercuartil; lo que permite detectar posibles tendencias en la distribución de la intensidad del entrenamiento. Esta gráfica nos permite observar las relaciones potenciales entre la carga de entrenamiento y la ocurrencia de lesiones, permitiendo evaluar si existen diferencias significativas en la distribución de *Training_Intensity* entre jugadores con mayor y menor predisposición a lesiones.

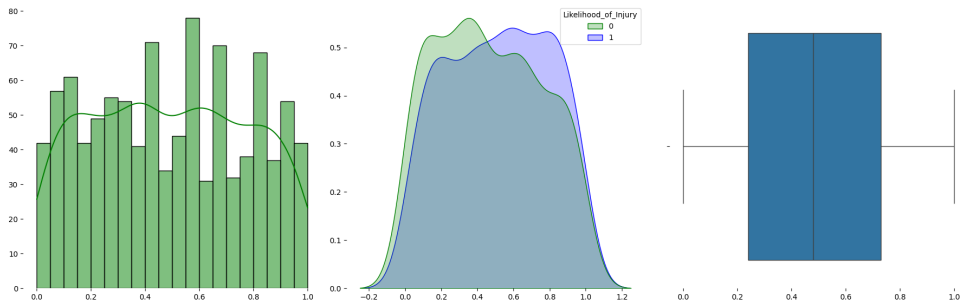


Figura 2.6: Distribución y análisis de la intensidad del entrenamiento en relación con la probabilidad de lesión.
Fuente: Los Autores.

- **Análisis del Tiempo de Recuperación y su Relación con la Probabilidad de Lesión**

  Para evaluar la variable *Recovery_Time* y su posible correlación con la probabilidad de lesión, se han generado unas gráficas como se muestran en la figura 2.7 compuesta por un histograma con estimación de densidad de núcleo (KDE), un gráfico de densidad condicionado a la probabilidad de lesión y un diagrama de caja y bigotes. El histograma muestra la

frecuencia de diferentes valores de tiempo de recuperación, permitiendo identificar tendencias en la duración de los periodos de recuperación de los jugadores. La estimación de densidad permite una comparación detallada de la distribución de *Recovery_Time* entre los jugadores con y sin historial de lesión, proporcionando una visión más continua de la variabilidad en los tiempos de recuperación. El diagrama de caja y bigotes ofrece información sobre la dispersión de los valores, destacando la mediana, el rango intercuartil y la presencia de valores atípicos. Esta gráfica es útil para detectar patrones en los tiempos de recuperación, y evaluar si existen diferencias significativas entre jugadores con distinta predisposición a lesiones.

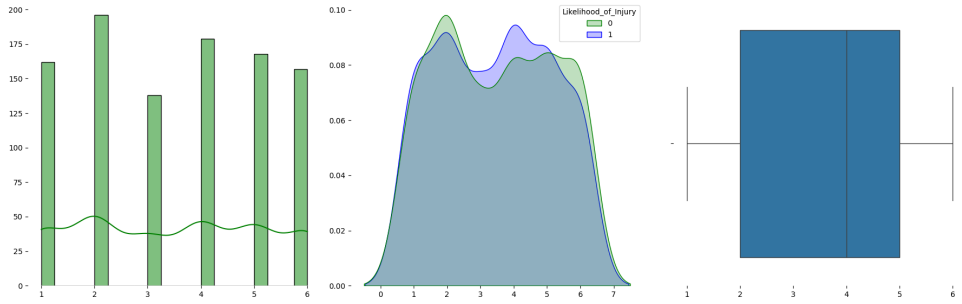


Figura 2.7: Distribución y análisis del tiempo de recuperación en relación con la probabilidad de lesión.
Fuente: Los Autores.

- **Distribución del Riesgo de Lesión**

  El análisis de la distribución de la variable *Likelihood_of_Injury* se presenta mediante un gráfico de pastel, en el que se muestra la proporción de jugadores clasificados con riesgo de lesión y aquellos sin riesgo. Como se observa en la Figura 2.8, la distribución está equilibrada de manera uniforme, con un 50 % de los jugadores perteneciendo a la categoría de riesgo (1) y el otro 50 % sin riesgo (0). Este balance es fundamental para garantizar un entrenamiento adecuado de los modelos de clasificación, evitando sesgos que podrían surgir en presencia de una distribución desbalanceada. Un *dataset* desbalanceado podría favorecer la clase mayoritaria, disminuyendo la capacidad del modelo para identificar de forma correcta los casos menos representados. El balanceo en la distribución nos indica que los datos han sido procesados o seleccionados de manera cuidadosa.

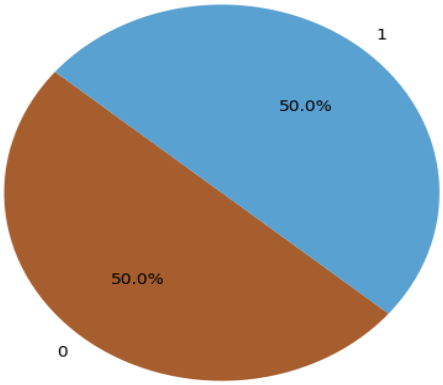


Figura 2.8: Distribución de la variable *Likelihood_of_Injury*.
Fuente: Los Autores.

2.2. Envío y recepción de datos en tiempo real

La conexión con ThingsBoard se llevó a cabo utilizando la biblioteca `requests` de Python, la cual ofrece una forma sencilla y eficiente de interactuar con APIs mediante solicitudes HTTP. Para enviar datos en tiempo real, seguimos los pasos detallados a continuación:

1. **Preparación de los datos**

   Los datos utilizados en este proyecto provinieron de un archivo Excel que contenía información detallada de jugadores, estos datos se procesaron y estructuraron en Python usando bibliotecas como `pandas`, y se almacenaron en un diccionario para enviarlos en formato JSON al servidor de ThingsBoard.

2. **Envío de datos mediante POST**

   Se utilizó el método `POST` para enviar los datos estructurados al *endpoint* de ThingsBoard. Este método permite enviar datos al servidor para su almacenamiento o procesamiento.

3. **Validación de la respuesta**

   La respuesta del servidor fue validada para asegurarnos de que los datos se enviaron de manera correcta. Un código de estado `200` indicó éxito, mientras que otros códigos revelaron problemas que se corrigieron durante el desarrollo.

```
THINGSBOARD_URL = "http://3.15.222.219:8080/api/v1/"
ACCESS_TOKEN = "C9Ua56nNqjVWT1a00H1h"
```

```
thingsboard_url = f"{THINGSBOARD_URL}{ACCESS_TOKEN}/telemetry"
thingsboard_headers = {"Content-Type": "application/json"}

try:
    response_tb = requests.post(
    thingsboard_url,
    headers=thingsboard_headers,
    json=datos_thingsboard_con_recomendaciones)
    if response_tb.status_code == 200:
        print(f"Datos enviados correctamente
        para el jugador {index+1}")
    else:
        print(f"Error al enviar datos a ThingsBoard:
        {response_tb.status_code},
        {response_tb.text}")
except requests.exceptions.RequestException as e:
    print(f"Error al conectar con ThingsBoard: {e}")
```

2.3. Configuración y conexión con ThingsBoard

La configuración inicial de ThingsBoard incluyó la creación de un dispositivo que representa a cada grupo de jugadores. Estos dispositivos se configuraron con claves de acceso únicas (`ACCESS_TOKEN`) para garantizar una comunicación segura. Los pasos clave fueron:

1. **Creación de dispositivos:**
    - Se accedió al panel de administración de ThingsBoard.

- Se creó un nuevo dispositivo para cada grupo de jugadores, los dispositivos en ThingsBoard permiten el envío de datos, el dispositivo que se utilizo tiene una conexión http y compatible para un sistema operativo Windows.
- Se generaron tokens de acceso únicos asociados a cada dispositivo.

2. **Definición de telemetría:**

   - Se definieron las métricas que los dispositivos reportarían, como edad, peso, altura, intensidad de entrenamiento y probabilidad de lesión.
   - Se estableció un esquema para organizar los datos enviados desde Python.

3. **Envio de datos a ThingsBoard:**

   Una vez establecidas las métricas y creación de los dispositivos, los datos procesados se envian en tiempo real utilizando los endpoints de ThingsBoard.

   Los datos tienen la estructura en formato Json para el envio.

4. **Creación de dashboards en ThingsBoard:** Para la visualización de datos, se implemento dos dashboards personalizados: uno basado en una línea de tiempo y otro en un HTML Value Card.

   - Línea del tiempo: La primera visualización es un gráfico que permite observar la evolución de la intensidad de entrenamiento y la probabilidad de lesión a lo largo del tiempo.
   - HTML Value Card: El segundo dashboard utiliza un widget HTML Value Card para mostrar los consejos generados por Gemini, basados en los datos analizados por el modelo de ML.

5. **Sincronización de Datos con Dashboards:**

   La sincronización entre los dispositivos y los dashboards se logró mediante el mecanismo de suscripción de telemetría de ThingsBoard. Cada vez que el dispositivo enviaba nuevos datos, los widgets del dashboard se actualizaban de manera automática.

Figura 2.9: Dashboards de ThingsBoard.
Fuente: Los Autores.

2.4. Implementación con la Api de Gemini.

La implementación con la API de gemini utiliza datos que se obtuvieron de ThingsBoard con el fin de generar recomendaciones personalizadas sobre entrenamiento y prevención de lesiones a través de la API.

Primero, se configura la conexión al modelo Gemini-1.5-flash mediante su API. Esta API utiliza un modelo de ML avanzado para generar texto basado en los datos proporcionados.

```
genai.configure(api_key="AIzaSyCBaD3InDd8--C12AihiR-y3P2vDly09U")
model = genai.GenerativeModel("gemini-1.5-flash")
```

El modelo está diseñado para actuar como un experto en deportología, proporcionando recomendaciones basadas en la probabilidad de lesión y otros datos fisiológicos y de entrenamiento. Para obtener recomendaciones toma los datos específicos de un jugador y genera texto detallado sobre cómo prevenir lesiones.

```
response = model.generate_content(f"""
    Actúa como un experto en deportología.
    En base a los siguientes resultados de la Probabilidad de
        ↪ lesión y con los datos obtenidos que son Edad del
        ↪ jugador, Peso del jugador, Altura del jugador,
        ↪ Lesiones previas, Intensidad de entrenamiento y
        ↪ Tiempo de recuperación, ¿qué recomendaciones me
        ↪ darías para evitar lesiones durante el
        ↪ entrenamiento?

    Edad del jugador: {datos['Player_Age']} años
    Peso del jugador: {datos['Player_Weight']} kg
    Altura del jugador: {datos['Player_Height']} cm
    Lesiones previas: {datos['Previous_Injuries']}
    Intensidad de entrenamiento: {datos['Training_Intensity']}
    iempo de recuperación: {datos['Recovery_Time']} días
    Probabilidad de lesión:
        ↪ {datos['Prediction_Likelihood_of_Injury']}

    Realiza la recomendación en máximo 3 items, de dos lineas
        ↪ máximo cada item.
    Y a demas pongo en un html.
    """)
```

Los datos de entrada se obtienen desde un archivo Excel llamado predicciones.xlsx, que contiene las características de los jugadores y las probabilidades de lesión generadas por un modelo de ML. Cada fila del archivo representa un jugador.

Los datos son iterados y enviados a ThingsBoard junto con las recomendaciones generadas por Gemini, esta recomendación es enviada a través de un dispositivo

configurado en ThingsBoard y la respuesta es mostrada en un dashboard llamado value card, el cual me permite mediante etiquetas HTML colocar el valor del campo.

Capítulo 3

Análisis de Resultados

En este capítulo se muestran los resultados obtenidos de la implementación y evaluación de los modelos de ML empleados para predecir el riesgo de lesiones en deportistas. El proceso de evaluación se realizó en diversas etapas, comenzando por la selección de las métricas adecuadas para medir el desempeño de los modelos, tales como *accuracy*, *precision*, y *recall*. A continuación, se llevó a cabo un análisis más detallado utilizando herramientas adicionales como las curvas características operativas del receptor (ROC, del inglés *Receiver Operating Characteristic*) y las matrices de confusión, para obtener una visión más completa del rendimiento de cada modelo. El uso de las curvas ROC permiten evaluar el rendimiento general de una prueba para compararlo con diferentes métodos de ML. De igual modo, se usaron matrices de confusión ya que permiten evaluar el rendimiento de la respuesta de clasificación de los modelos de interés.

3.1. Evaluación del Desempeño de los Modelos

Para evaluar la capacidad predictiva de los modelos de ML implementados, se utilizaron métricas estándar en clasificación: *accuracy*, *precision* y *recall*.

- **Métricas de Evaluación** La métrica *accuracy* representa la proporción de predicciones correctas sobre el total de observaciones, proporcionando una visión general del desempeño del modelo. La *precision* mide las instancias que se clasificaron de manera correcta como positivas respecto al total de

predicciones positivas realizadas, mientras que el *recall* evalúa la proporción de casos positivos que fueron identificados de manera acertada por el modelo.

- **Resultados de la Evaluación**
  Cada modelo fue entrenado utilizando el conjunto de datos de entrenamiento y evaluado con los datos de prueba. En la Tabla 3.1 se presentan los resultados obtenidos:

Tabla 3.1: Evaluación del rendimiento de los modelos de ML.

Modelo	Accuracy	Precision	Recall
LGBMClassifier	0.57	0.6739	0.5254
AdaBoostClassifier	0.63	0.7037	0.6441
ExtraTreesClassifier	0.54	0.6275	0.5424
NuSVC	0.50	0.5882	0.5085
ExtraTreeClassifier	0.58	0.6393	0.6610
SVM	0.59	0.7143	0.5085

3.2. Evaluación del Desempeño mediante Curvas ROC

Para analizar el desempeño de los modelos utilizados en la predicción del riesgo de lesión, se empleó la Curva ROC y el cálculo del área bajo la curva AUC (del inglés Area Under the Curve). Este análisis permite evaluar la habilidad para diferenciar entre atletas propensos a sufrir lesiones y aquellos que no lo están, se obtiene al graficar la tasa de verdaderos positivos (True Positive Rate, TPR) frente a la tasa de falsos positivos (False Positive Rate, FPR) para distintos umbrales de decisión. El AUC representa la capacidad general del modelo para realizar una clasificación correcta:

- Un AUC cercano a 1 indica una alta capacidad discriminativa del modelo.
- Un AUC cercano a 0.5 sugiere que el modelo no es mejor que una clasificación aleatoria.
- Un AUC inferior a 0.5 implicaría que el modelo tiene un desempeño peor que el azar, indicando una inversión en la clasificación.

Se ajustaron los modelos a los datos de entrenamiento y se generaron predicciones probabilísticas sobre el conjunto de prueba, es con esta información se calcularon las Curva **ROC** y el **AUC** para cada modelo.

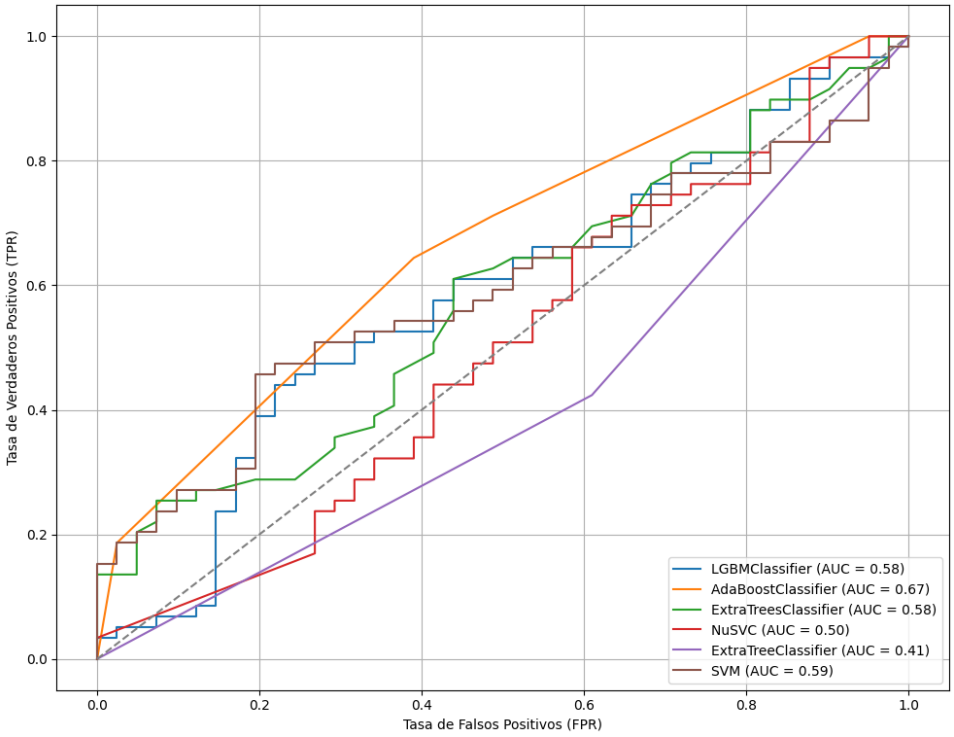


Figura 3.1: Curvas ROC obtenidas para los modelos evaluados.
Fuente: Los Autores.

Como se puede ver en la figura 3.1, los modelos con curvas por encima de la línea punteada (AdaBoostClassifier, LGBMClassifier y SVM) son los más prometedores, con oportunidad de optimización Los modelos con curvas por debajo de la línea punteada (ExtraTreeClassifier) deben revisarse, ya que su rendimiento es inferior al de un clasificador aleatorio; y los modelos con curvas que coinciden con la línea punteada (NusVC) no son útiles para la tarea de clasificación en su estado actual.

La Tabla 3.2 muestra los valores de **AUC** obtenidos para cada modelo, proporcionando una comparación cuantitativa de su desempeño.

La tabla 3.2 muestra al modelo AdaBoostClassifier como el de mayor valor de AUC (0.67), lo que indica una capacidad moderada para distinguir entre clases positivas y negativas, los modelos LGBMClassifier y ExtraTreesClassifier presentaron un AUC idéntico (0.58), mostrando un rendimiento ligeramente superior al azar (AUC = 0.5), podría atribuirse a una sensibilidad limitada ante patrones complejos en

Tabla 3.2: Valores de AUC obtenidos para cada modelo.

Modelo	AUC
LGBMClassifier	0.58
AdaBoostClassifier	0.67
ExtraTreesClassifier	0.59
NuSVC	0.50
ExtraTreeClassifier	0.41
SVM	0.59

los datos o a la necesidad de ajustar hiperparámetros críticos, como la profundidad de los árboles o la tasa de aprendizaje. El algoritmo SVM registró un AUC de 0.59, similar a los anteriores, lo que refleja un desempeño modesto.

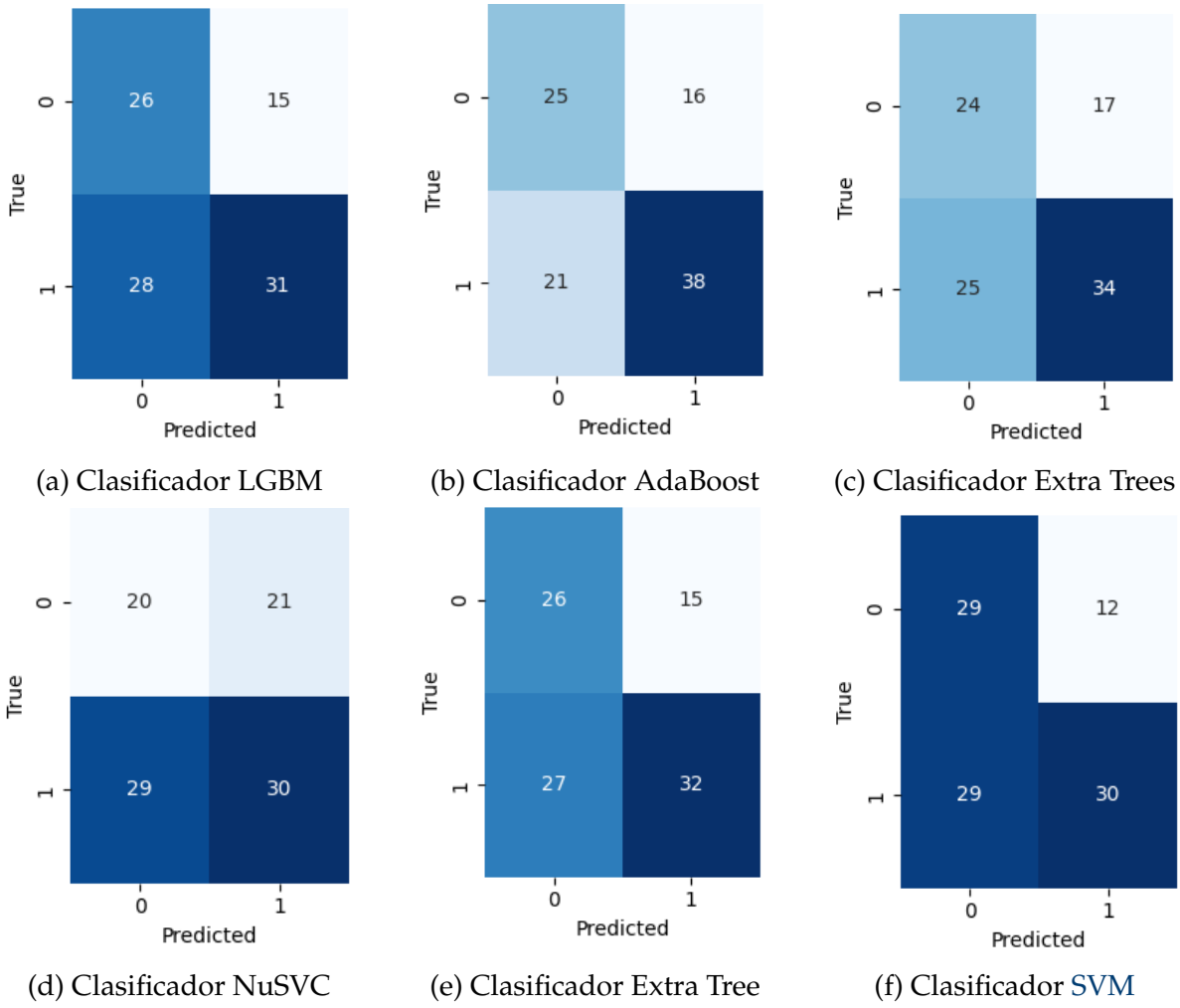
3.3. Evaluación mediante Matrices de Confusión

Para analizar el desempeño de los modelos de clasificación, se utilizaron matrices de confusión ya que permiten visualizar la distribución de predicciones correctas e incorrectas Así, al proporcionar una descomposición detallada de los aciertos y errores en las predicciones facilita identificar entre verdaderos positivos, falsos positivos, verdaderos negativos y falsos negativos.

En la Figura 3.2 se presentan las matrices de confusión obtenidas para cada modelo y se observa que algunos modelos presentan un mejor desempeño en la clasificación que otros. Esto se debe al hecho de manejar de manera más eficiente volúmenes de datos, incluir técnicas para evitar sobreajuste, introducir aleatoriedad en umbrales de división y tener mejor desempeño ante datos etiquetados.

LGBMClassifier en 3.2a, presentó 26 verdaderos negativos y 31 verdaderos positivos, con 15 falsos positivos y 28 falsos negativos, entonces el modelo tiene una capacidad moderada de clasificación, pero presenta un número significativo de falsos negativos, lo que podría afectar su capacidad predictiva en la identificación de eventos positivos.

AdaBoostClassifier en 3.2b, logró 25 verdaderos negativos y 38 verdaderos positivos, con 16 falsos positivos y 21 falsos negativos; el modelo muestra un mejor



(a) Clasificador LGBM (b) Clasificador AdaBoost (c) Clasificador Extra Trees

(d) Clasificador NuSVC (e) Clasificador Extra Tree (f) Clasificador SVM

Figura 3.2: Matriz de confusión para distintos clasificadores.

balance entre ambas clases, logrando reducir la cantidad de falsos negativos en comparación con LGBMClassifier, muestra un mejor desempeño en la clasificación de eventos positivos.

ExtraTreesClassifier en 3.2c, clasificó de forma correcta 24 casos negativos y 34 casos positivos, con 17 falsos positivos y 25 falsos negativos. Este modelo mantiene una precisión estable, aunque con una reducción en la tasa de verdaderos positivos, lo que puede impactar en su clasificación.

NuSVC en 3.2d, presentó un desempeño inferior en términos de clasificación con 20 verdaderos negativos y 30 verdaderos positivos, con 21 falsos positivos y 29 falsos negativos. La distribución de los errores en ambas clases indica que el modelo no logra una separación eficiente entre las categorías, aumentando la probabilidad de error en la clasificación.

ExtraTreeClassifier en 3.2e, obtuvo 26 verdaderos negativos y 32 verdaderos positivos, con 15 falsos positivos y 27 falsos negativos. Su rendimiento es comparable al de LGBMClassifier, demostrando una mejor capacidad para identificar eventos positivos, aunque presenta una cantidad considerable de falsos negativos.

SVM en 3.2f mostró un desempeño aceptable con 29 verdaderos negativos y 30 verdaderos positivos, mientras que los errores corresponden a 12 falsos positivos y 29 falsos negativos. Este modelo la clasificación de eventos negativos es buena, pero la cantidad de falsos negativos podría requerir ajustes en los hiperparámetros para mejorar la sensibilidad.

En términos generales, **AdaBoostClassifier y ExtraTreesClassifier** presentan un mejor equilibrio entre falsos positivos y falsos negativos, lo que sugiere como una buena opción para la tarea de clasificación. **LGBMClassifier y SVM** ofrecen un rendimiento aceptable.

3.4. Evaluación por métrica F1

Para determinar el mejor modelo, se emplea un análisis cuantitativo de las métricas clave de clasificación: Recall y Precision, como se explicó en la sección 1.2.1 se utilizó la siguiente fórmula de puntuación F1.

```
F1_score = ((2 * recall * precision)/(recall + precision))
```

Este cálculo permite integrar un valor único el cual representa el desempeño general de cada modelo.

```
Mejor modelo: AdaBoostClassifier, con puntuación ponderada: 0.6095
Ranking de Modelos (mejor a peor):
1. AdaBoostClassifier => Puntuación ponderada: 0.6095
2. SVM => Puntuación ponderada: 0.5941
3. LGBMClassifier => Puntuación ponderada: 0.5905
4. ExtraTreesClassifier => Puntuación ponderada: 0.5586
5. NuSVC => Puntuación ponderada: 0.5455
```

```
6. ExtraTreeClassifier => Puntuación ponderada: 0.5143

Mejor modelo: AdaBoostClassifier, con puntuación ponderada: 0.6095
```

Capítulo 4

Conclusiones, Recomendaciones y Trabajos Futuros

Conclusiones

El presente trabajo ha desarrollado un sistema de monitoreo basado en IA para la predicción del riesgo de lesión, decisión que se ha tomado tras la tendencia actual analizada en el estado del arte dentro de la medicina deportiva en las tres aplicaciones más comunes: (1) predicción del rendimiento, (2) prevención de lesiones y (3) reconocimiento de patrones. Los modelos mostraron un desempeño variable en términos de precisión (*precision*), exhaustividad (*recall*) y exactitud (*accuracy*), donde AdaBoost, SVM y LGBM destacaron como los 3 mejores para identificar patrones de riesgo y predicción de lesiones a partir de variables físicas y cargas de entrenamiento. El modelo *AdaBoostClassifier* obtuvo el mejor balance entre estas métricas, lo que sugiere su idoneidad para la tarea de predicción en el contexto de un *dataset* de carácter limitado. Sin embargo, otros modelos como *LGBMClassifier* y SVM presentaron valores competitivos, lo que indica que su desempeño puede mejorarse con técnicas de optimización de ajustes finos de hiperparámetros como la tasa de aprendizaje , la profundidad máxima , el número de estimadores , y la fracción de submuestreo en el caso de LGBM que equilibran el sesgo-varianza, evitando el sobreajuste; o en el caso de hiperparámetros SVM como el tipo de kernel (lineal, polinomial o RBF), el parámetro de regularización (C), gamma (para kernels

no lineales) y la tolerancia a errores (tol) que ayudan a la capacidad para modelar fronteras de decisión.

El análisis de correlación permitió identificar que ciertas variables, como la intensidad del entrenamiento y el tiempo de recuperación, tienen un impacto significativo en la predicción del riesgo de lesión. Las curvas ROC y las matrices de confusión evidenciaron la capacidad discriminativa de los modelos para diferenciar entre deportistas en riesgo y aquellos sin riesgo significativo, mostrando que los modelos de *boosting* y máquinas de soporte vectorial lograron una mejor clasificación en términos de la relación entre verdaderos positivos y falsos positivos.

La implementación de un dashboard basado en IA para la predicción del riesgo de lesión es viable y aporta valor en la prevención y gestión de lesiones deportivas.

La integración del dashboard con *ThingsBoard* permitió la visualización y monitoreo en tiempo real de las variables fisiológicas y físicas de los deportistas, facilitando la toma de decisiones basada en los modelos de ML implementados. La conexión establecida mediante la API de ThingsBoard y el uso de un script en *Python* garantizando el envío de datos de manera estructurada, permitiendo que las métricas obtenidas sean representadas gráficamente en la interfaz del dashboard.

El modelo de IA seleccionado fue incorporado en el sistema mediante la API de *Gemini*, proporcionando recomendaciones para la prevención de lesiones. Dichas recomendaciones fueron generadas a partir de un análisis de factores como edad, peso, altura, historial de lesiones, intensidad del entrenamiento y tiempo de recuperación. La implementación del algoritmo verifica que los usuarios puedan recibir sugerencias personalizadas para mejorar su rendimiento deportivo y minimizar riesgos de lesión.

Trabajos Futuros

El análisis sugiere que futuras versiones del modelo podrían beneficiarse de la incorporación de nuevas características fisiológicas y biomecánicas, tales como la variabilidad de la frecuencia cardíaca o la carga externa del entrenamiento. Además, se detectó un margen de mejora en la reducción de falsos negativos, lo que podría

mejorar aún más si variables físicas, fisiológicas, psicológicas y de entorno se añaden al dataset, permitiendo un enfoque holístico en la predicción de lesiones.

Uno de los desafíos técnicos identificados en el desarrollo del sistema fue la dificultad de conseguir conjuntos de datos de entrenamiento etiquetados de alta calidad debido a problemas de seguridad y privacidad asociados con el manejo de datos personales, así como las restricciones de conexión con la API de Gemini. Esto resalta la necesidad de contar con un dataset de calidad que contemple variables de alta correlación con la variable objetivo y la importancia de implementar mecanismos de control en la comunicación con servicios externos para evitar bloqueos en el procesamiento de datos.

Este estudio abre la puerta a futuras investigaciones en la intersección entre IA y ciencias del deporte, permitiendo la optimización de metodologías de análisis y prevención de lesiones en diversas disciplinas deportivas.

Recomendaciones

Con base en los resultados obtenidos en la evaluación de los modelos de ML para la predicción de lesiones en deportistas, se sugieren las siguientes recomendaciones para mejorar la precisión del sistema, optimizar la implementación en entornos reales y evitar posibles limitaciones identificadas.

Primero es importante mencionar que encontrar patrones en grandes volúmenes de datos no garantiza que esos patrones sean útiles o correctos para explicar fenómenos complejos, como el rendimiento y desarrollo de los atletas, esto significa que los investigadores deben contextualizar los datos, considerar factores externos y aplicar principios científicos sólidos para evitar interpretaciones o conclusiones erróneas. Sin esta interpretación rigurosa, los patrones identificados pueden ser coincidencias, reflejar sesgos en los datos o carecer de relevancia en el mundo real.

Segundo, se recomienda la recopilación de un mayor número de deportistas, así como adicionar al menos las siguientes variables que abarcan tanto aspectos físicos como psicológicos del atleta: Distancia recorrida, alta velocidad, frecuencia

cardíaca, intensidad de velocidad, carga de estrés dinámica, carga a baja velocidad, impacto, aceleración, desaceleración, *sprints*, índices de fatiga, gasto energético, carga de entrenamiento e índices del cuestionario *hopper* [50].

Debido a la restricción por las leyes de protección de datos, para validar el correcto funcionamiento del modelo en entornos reales, sería ideal su implementación en clubes deportivos y centros de alto rendimiento, donde se pueden obtener *datasets* de mejor calidad que reflejen una mayor correlación con la variable objetivo, se sugieren registros de rendimiento de los atletas, historias clínicas y sobre todo datos de tecnología *wearable*. La evaluación del sistema en condiciones de uso reales y controlada permitirá identificar posibles limitaciones y ajustar el modelo a las necesidades específicas de los entrenadores y médicos deportivos; o en su defecto responder preguntas sobre cómo manejar datos incompletos y, al mismo tiempo, garantizar la privacidad del paciente utilizando la SDG como una herramienta para cerrar estas brechas bajo diferentes enfoques, por ejemplos la revisión sistemática presentada en [51] donde se sugieren métodos de referencia, modelos estadísticos, modelos de aprendizaje automático ML, enfoques de aprendizaje profundo por redes neuronales, redes generativas antagónicas (del inglés, Generative Adversarial Networks GAN), entre otros.

Se recomienda la exploración de técnicas de aprendizaje continuo (*Continual Learning*) que permitan actualizar el modelo de forma dinámica a medida que se dispone de nuevos datos sin necesidad de un re-entrenamiento completo, pues como se señala en [52] cada conjunto de datos necesita de un diseño de algoritmos y modelos de minería de datos específico; esto aseguraría que el sistema pueda adaptarse a cambios en las variables de entrenamiento y mejorar su capacidad de predicción a lo largo del tiempo.

La conexión con la API de Gemini presenta restricciones en cuanto a la frecuencia de solicitudes, lo que puede ser interpretado como un ataque automatizado, si no se establece un temporizador entre las llamadas a la API, se recomienda la implementación de retardos controlados (*timeouts*) en los procesos de comunicación con la API para evitar bloqueos y garantizar la estabilidad del sistema.

Glosario

AUC área bajo la curva.

DL Aprenizaje profundo o Deep Learning.

IA Inteligencia Artificial – Artifical Intelligence.

IoT Internet de las cosas.

LGBM Light Gradient Boosted Machine.

ML Maquina de aprendizaje o Machine Learning.

Organización Mundial de la Salud Organización Mundial de la Salud.

RBF Funciones de Base Radial del inglés Radial Basis Functions.

ROC Receiver Operating Characteristic.

RPE Índice de esfuerzo percibido o Rate of Perceived Exertion.

SDG Generación de datos sintéticos..

SVM Support Vector Machine - Máquina de vectores de soporte.

Referencias

[1] J. Clear, *Atomic Habits: An Easy & Proven Way to Build Good Habits & Break Bad Ones*. New York: Avery, 2018.

[2] S. Mohsen, A. Zekry, K. Youssef y M. Abouelatta, «A Self-powered Wearable Wireless Sensor System Powered by a Hybrid Energy Harvester for Healthcare Applications,» *Wireless Personal Communications*, vol. 116, n.º 4, págs. 3143-3164, 2021. DOI: 10.1007/s11277-020-07840-y. dirección: https://doi.org/10.1007/s11277-020-07840-y.

[3] D. Jiang, B. Shi, H. Ouyang, Y. Fan, Z. L. Wang y Z. Li, «Emerging Implantable Energy Harvesters and Self-Powered Implantable Medical Electronics,» *ACS Nano*, vol. 14, n.º 6, págs. 6436-6448, 2020. DOI: 10.1021/acsnano.9b08268. dirección: https://doi.org/10.1021/acsnano.9b08268.

[4] I. Awolusi, C. Nnaji, E. Marks y M. Hallowell, «Enhancing Construction Safety Monitoring through the Application of Internet of Things and Wearable Sensing Devices: A Review,» en *Computing in Civil Engineering 2019*, págs. 530-538. DOI: 10.1061/9780784482438.067. dirección: https://ascelibrary.org/doi/abs/10.1061/9780784482438.067.

[5] W. Liu, Z. Long, G. Yang y L. Xing, «A Self-Powered Wearable Motion Sensor for Monitoring Volleyball Skill and Building Big Sports Data,» *Biosensors*, vol. 12, n.º 2, 2022, ISSN: 2079-6374. DOI: 10.3390/bios12020060. dirección: https://www.mdpi.com/2079-6374/12/2/60.

[6] L. Manjakkal, S. Mitra, Y. R. Petillot et al., «Connected Sensors, Innovative Sensor Deployment, and Intelligent Data Analysis for Online Water Quality Monitoring,» *IEEE Internet of Things Journal*, vol. 8, n.º 18, págs. 13 805-13 824, 2021. DOI: 10.1109/JIOT.2021.3081772. dirección: https://ieeexplore.ieee.org/document/9435802.

[7] T. G. Stavropoulos, A. Papastergiou, L. Mpaltadoros, S. Nikolopoulos e I. Kompatsiaris, «IoT Wearable Sensors and Devices in Elderly Care: A Literature Review,» *Sensors*, vol. 20, n.º 10, 2020, ISSN: 1424-8220. DOI: 10.3390/s20102826. dirección: https://www.mdpi.com/1424-8220/20/10/2826.

[8] K. K. B. Peetoom, M. A. S. Lexis, M. Joore, C. D. Dirksen y L. P. D. Witte, «Literature review on monitoring technologies and their outcomes in independently living elderly people,» *Disability and Rehabilitation: Assistive Technology*, vol. 10, n.º 4, págs. 271-294, 2015. DOI: 10.3109/17483107.2014.961179. dirección: https://doi.org/10.3109/17483107.2014.961179.

[9] D. Nazarevich, *Tendencias de big data 2024: Navegando por el futuro de la tecnología de datos*, Accessed on September 24, 2024, 2024. dirección: https://innowise.com/es/blog/big-data-trends-2024/.

[10] G. Liu, Y. Luo, O. Schulte y T. Kharrat, «Deep soccer analytics: learning an action-value function for evaluating soccer players,» *Data Mining and Knowledge Discovery*, vol. 32, n.º 5, págs. 1531-1559, 2020. DOI: 10.1007/s10618-020-00705-9.

[11] Y. Córdova Viera, J. Martínez Borrego y E. Córdova Viera, «Propuesta de metodología para el diseño de dashboard,» *Revista Cubana de Transformación Digital*, vol. 2, n.º 3, oct. de 2021. DOI: 10.5281/zenodo.5545998.

[12] J. Amissah, O. Abdel-Rahim, D. Mansour et al., «Developing a three stage coordinated approach to enhance efficiency and reliability of virtual power plants,» *Scientific Reports*, vol. 14, pág. 13 105, 2024. DOI: 10.1038/s41598-024-63668-7.

[13] H. Wang, «NBA Game Analysis based on Big Data,» *Highlights in Science, Engineering and Technology*, vol. 24, págs. 54-57, 2022. DOI: 10.54097/hset.v24i.3885. dirección: https://drpress.org/ojs/index.php/HSET/article/view/3885.

[14] E. E. Cuenca Pauta y B. A. Corella Parra, «Using big data techniques to measure the performance of professional basketball teams,» Tesis de mtría., Universidad de Investigación de Tecnología Experimental Yachay, 2021.

[15] F. E. Arcos Coronel y J. S. Quezada Patiño, «Desarrollo de un sistema de monitoreo de frecuencia cardiaca y variables físicas para medir el desempeño en sesiones de entrenamiento de futbolistas usando tecnologías de comunicación inalámbrica.,» Tesis de mtría., Universidad Politécnica Salesiana, 2023.

[16] Organización Mundial de la Salud (OMS), *Actividad física*, Consultado el 24 de septiembre de 2023, 2023. dirección: https://www.who.int/es/news-room/fact-sheets/detail/physical-activity.

[17] Barça Innovation Hub, *¿Cuánto se lesiona un jugador profesional? Un análisis de la epidemiología de las lesiones en el fútbol*, Consultado el 24 de septiembre de 2023, 2023. dirección: https://barcainnovationhub.fcbarcelona.com/es/blog/cuanto-se-lesiona-un-jugador-profesional-un-analisis-de-la-epidemiologia-de-las-lesiones-en-el-futbol/.

[18] Mordor Intelligence, *Wearable Devices in Sports Market - Growth, Trends, COVID-19 Impact, and Forecasts (2023 - 2028)*, Consultado el 24 de septiembre de 2023, 2023. dirección: https://www.mordorintelligence.com/es/industry-reports/wearable-devices-in-sports-market.

[19] A. Vargas-Pacheco y L. E. Correa-López, «El ejercicio como protagonista en la plasticidad muscular y en el músculo como un órgano endocrino: Implicaciones en las enfermedades crónicas,» es, *Revista de la Facultad de Medicina Humana*, vol. 22, págs. 181-192, ene. de 2022, ISSN: 2308-0531. dirección: http://www.scielo.org.pe/scielo.php?script=sci_arttext&pid=S2308-05312022000100181&nrm=iso.

[20] K. A. Q. Veloz, C. I. C. Llumiquinga y E. R. T. Iza, «La transformación digital en el deporte: El impacto de las TICs en la mejora del rendimiento deportivo y la experiencia del usuario: Digital transformation in sport: The Impact of ICTs on improving sports performance and user experience,» *LATAM Revista Latinoamericana de Ciencias Sociales y Humanidades*, vol. 5, n.º 4, págs. 1145-1154, 2024. DOI: 10.56712/latam.v5i4.2321.

[21] N. Terrados, J. Calleja-González y X. Schelling, «Bases fisiológicas comunes para deportes de equipo,» *Revista Andaluza de Medicina del Deporte*, 2011, ISSN: 1888-7546. dirección: https://www.elsevier.es/es-revista-revista-andaluza-medicina-del-deporte-284-articulo-bases-fisiologicas-comunes-deportes-equipo-X1888754611213192.

[22] C. Granados, M. Izquierdo, J. Ibáñez, M. Ruesta y E. M. Gorostiaga, «Effects of an Entire Season on Physical Fitness in Elite Female Handball Players,» *Medicine & Science in Sports & Exercise*, vol. 40, n.º 2, págs. 351-361, feb. de 2008. DOI: 10.1249/mss.0b013e31815b4905.

[23] I. Loturco, L. A. Pereira, V. P. Reis et al., «Power training in elite young soccer players: Effects of using loads above or below the optimum power zone,» *Journal of Sports Sciences*, vol. 38, n.º 11-12, págs. 1416-1422, jun. de 2020, Epub 2019 Aug 7. DOI: 10.1080/02640414.2019.1651614.

[24] L. Fernández-Galván, P. Jimenez-Reyes, V. Cuadrado y A. Casado, «Sprint Performance and Mechanical Force-Velocity Profile among Different Maturational Stages in Young Soccer Players,» *International Journal of Environmental Research and Public Health*, vol. 19, ene. de 2022. DOI: 10.3390/ijerph19031412.

[25] E. J. Bradshaw y P. Le Rossignol, «Anthropometric and biomechanical field measures of floor and vault ability in 8 to 14 year old talent-selected gymnasts,» *Sports Biomechanics*, vol. 3, n.º 2, págs. 249-262, jul. de 2004. DOI: 10.1080/14763140408522844.

[26] F. Bianchi, T. Facchinetti y P. Zuccolotto, «Role revolution: Towards a new meaning of positions in basketball,» *Electronic Journal of Applied Statistical Analysis*, vol. 10, págs. 712-734, ene. de 2017. DOI: 10.1285/i20705948v10n3p712.

[27] Z. Bai y X. Bai, «Sports Big Data: Management, Analysis, Applications, and Challenges,» *Complexity*, vol. 2021, n.º 1, pág. 6 676 297, 2021. DOI: 10.1155/2021/6676297. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1155/2021/6676297. dirección: https://onlinelibrary.wiley.com/doi/abs/10.1155/2021/6676297.

[28] H. Murtaza, M. Ahmed, N. F. Khan, G. Murtaza, S. Zafar y A. Bano, «Synthetic data generation: State of the art in health care domain,» *Computer Science Review*, vol. 48, pág. 100 546, 2023, ISSN: 1574-0137. DOI: 10.1016/j.cosrev.2023.100546. dirección: https://www.sciencedirect.com/science/article/pii/S1574013723000138.

[29] IEAD - Instituto Europeo de Alta Dirección, *Big Data: El nuevo petróleo del siglo XXI (y cómo no ahogarte en él)*, 2024. dirección: https://iead.es/big-data-el-nuevo-petroleo-del-siglo-xxi-y-como-evitar-ahogarte-en-el/.

[30] J. G. Claudino, D. d. O. Capanema, T. V. de Souza, J. C. Serrão, A. C. M. Pereira y G. P. Nassis, «Current Approaches to the Use of Artificial Intelligence for Injury Risk Assessment and Performance Prediction in Team Sports: a Systematic Review,» *Sports Medicine-Open*, vol. 5, n.º 1, págs. 1-9, 2019. DOI: 10.1186/s40798-019-0202-3. dirección: https://doi.org/10.1186/s40798-019-0202-3.

[31] S. E. Woo, L. Tay, A. T. Jebb, M. T. Ford y M. L. Kern, «Big Data for Enhancing Measurement Quality,» en *Big Data in Psychological Research*, S. E. Woo, L. Tay y R. W. Proctor, eds., American Psychological Association, 2020, págs. 59-85. DOI: 10.1037/0000193-004.

[32] S. J. Russell y P. Norvig, *Artificial Intelligence: A Modern Approach*, 4th. Pearson, 2020, ISBN: 9780134610993.

[33] F. Thabtah, L. Zhang y N. Abdelhamid, «NBA Game Result Prediction Using Feature Analysis and Machine Learning,» *Annals of Data Science*, vol. 6, págs. 103-116, 2019. DOI: 10.1007/s40745-018-00189-x.

[34] D. A. Crawford y T. Davis, «Using K-means clustering to individualize training for collegiate basketball athletes following pre-season performance testing,» *International Journal of Exercise Science: Conference Proceedings*, vol. 11, n.º 11, Article 96, 2024. dirección: https://digitalcommons.wku.edu/ijesab/vol11/iss11/96.

[35] J. E. van Engelen y H. H. Hoos, «A survey on semi-supervised learning,» *Machine Learning*, vol. 109, n.º 2, págs. 373-440, 2020. DOI: 10.1007/s10994-019-05855-6.

[36] S. Chidambaram, Y. Maheswaran, K. Patel et al., «Using Artificial Intelligence-Enhanced Sensing and Wearable Technology in Sports Medicine and Performance Optimisation,» *Sensors*, vol. 22, n.º 18, pág. 6920, 2022. DOI: 10.3390/s22186920. dirección: https://doi.org/10.3390/s22186920.

[37] W. Li, «A Big Data Approach to Forecast Injuries in Professional Sports Using Support Vector Machine,» *Mobile Networks and Applications*, 2024. DOI: 10.1007/s11036-024-02377-x. dirección: https://doi-org.ecups.idm.oclc.org/10.1007/s11036-024-02377-x.

[38] A. Wong, E. Li, H. Le, G. Bhangu y S. Bhatia, «A predictive analytics framework for forecasting soccer match outcomes using machine learning models,» *Decision Analytics Journal*, vol. 14, pág. 100 537, 2025, ISSN: 2772-6622. DOI: 10.1016/j.dajour.2024.100537. dirección: https://www.sciencedirect.com/science/article/pii/S2772662224001413.

[39] Y. Kong y Z. Duan, «Boxing behavior recognition based on artificial intelligence convolutional neural network with sports psychology assistant,» *Scientific Reports*, vol. 14, n.º 7640, 2024. DOI: 10.1038/s41598-024-58518-5. dirección: https://www.nature.com/articles/s41598-024-58518-5.

[40] L. Guo, «Analysis and prediction of athlete's anxiety state based on artificial intelligence,» *PeerJ. Computer science*, vol. 9, 2023. DOI: 10.7717/peerj-cs.1322. dirección: https://pubmed.ncbi.nlm.nih.gov/37346592/.

[41] M. Meyer, C. Kandathil, S. Davis et al., «Evaluation of Rhinoplasty Information from ChatGPT, Gemini, and Claude for Readability and Accuracy,» *Aesthetic Plastic Surgery*, vol. 48, n.º 4, págs. 1165-1174, 2024, Received: 19 June 2024, Accepted: 23 August 2024, Published: 16 September 2024. DOI: 10.1007/s00266-024-04343-0. dirección: https://doi.org/10.1007/s00266-024-04343-0.

[42] G. Pillitteri, A. Rossi, C. Simonelli, I. Leale, V. Giustino y G. Battaglia, «Association between internal load responses and recovery ability in U19 professional soccer players: A machine learning approach,» *Heliyon*, vol. 9, n.º 4, págs. 2405-8440, 2023. DOI: 10.1016/j.heliyon.2023.e15454. dirección: https://doi.org/10.1016/j.heliyon.2023.e15454.

[43] D. Tang, «Systematic training of table tennis players' physical performance based on artificial intelligence technology and data fusion of sensing devices,» *SLAS technology*, vol. 29, n.º 4, 2024. DOI: 10.1016/j.slast.2024.100151. dirección: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85204416695&doi=10.1016%2fj.slast.2024.100151&partnerID=40&md5=9c3ef090192e0c2f07f1a255ddb74ee1.

[44] J. Fang y T. Xiang, «Medical Decision Support for Football Players Based on Machine Learning Historical Injury Data,» *Revista Internacional de Medicina y Ciencias de la Actividad Física y el Deporte*, vol. 24, n.º 96, págs. 479-489, 2024. DOI: 10.15366/rimcafd2024.96.029. dirección: https://dialnet.unirioja.es/servlet/articulo?codigo=9844992.

[45] L. Zhang, «Design and Adjustment of Optimizing Athletes' Training Programs Using Machine Learning Algorithms,» *Journal of Electrical Systems*, vol. 20, págs. 2014-2024, abr. de 2024. DOI: 10.52783/jes.3116.

[46] H. Tian, Y. Zhang y R. Ding, «Based on the LGBM model, a dynamic analysis was conducted to study the momentum of athletes in tennis matches,» *Transactions on Computer Science and Intelligent Systems Research*, vol. 6, págs. 229-237, oct. de 2024. DOI: 10.62051/f4c1pq70.

[47] G. Morciano, A. Zingoni y G. Calabrò, «Optimization and comparison of machine learning algorithms for the prediction of the performance of football players,» *Neural Comput. & Applic.*, vol. 36, págs. 19 653-19 666, 2024. DOI: 10.1007/s00521-024-10260-9. dirección: https://doi.org/10.1007/s00521-024-10260-9.

[48] MrSimple07, *Injury Prediction Dataset*, Accedido el 2 febrero del 2025, Kaggle, 2023. dirección: https://www.kaggle.com/datasets/mrsimple07/injury-prediction-dataset.

[49] D. D. Lab, *Feature Engineering*, https://domino.ai/data-science-dictionary/feature-engineering, Consultado el 2023-10-15, 2023. dirección: https://domino.ai/data-science-dictionary/feature-engineering.

[50] S. L. Hooper, L. T. Mackinnon, A. Howard, R. D. Gordon y A. W. Bachmann, «Markers for monitoring overtraining and recovery,» *Medicine and Science in Sports and Exercise*, vol. 27, n.º 1, págs. 106-112, ene. de 1995, ISSN: 0195-9131. DOI: 10.1249/00005768-199501000-00019. dirección: https://pubmed.ncbi.nlm.nih.gov/7898325/.

[51] M. Hernandez, G. Epelde, A. Alberdi, R. Cilla y D. Rankin, «Synthetic data generation for tabular health records: A systematic review,» *Neurocomputing*, vol. 493, págs. 28-45, 2022, ISSN: 0925-2312. DOI: 10.1016/j.neucom.2022.04.053. dirección: https://www.sciencedirect.com/science/article/pii/S0925231222004349.

[52] J. Cui, H. Du y X. Wu, «Data analysis of physical recovery and injury prevention in sports teaching based on wearable devices,» *Preventive Medicine*, vol. 173, pág. 107 589, 2023, ISSN: 0091-7435. DOI: 10.1016/j.ypmed.2023.107589. dirección: https://www.sciencedirect.com/science/article/pii/S009174352300169X.