

PONTIFICIA UNIVERSIDAD CATÓLICA DEL ECUADOR
FACULTAD DE CIENCIAS EXACTAS, NATURALES Y AMBIENTALES
CARRERA DE CIENCIA DE DATOS

MONOGRAFÍA

Nombre del estudiante: André Sebastián Quezada Guerron

Nombre del tutor: Mgtr. Kevin Ricardo Rojas Satian

Título de monografía: Predicción de Riesgo de Lesiones Deportivas mediante Algoritmos
de Aprendizaje Supervisado

Área en la que se inscribe la investigación: Aprendizaje Automático (Machine Learning)

Quito, 2026

CERTIFICACIÓN

El Mgtr. Kevin Ricardo Rojas Satian certifica que el trabajo de monografía titulado "Predicción de riesgo de lesiones deportivas mediante algoritmos de aprendizaje supervisado", realizado por André Sebastián Quezada Guerron, estudiante de la Carrera de Ciencia de Datos, ha sido concluido satisfactoriamente y cumple con los requisitos establecidos por la Facultad de Ciencias Exactas, Naturales y Ambientales de la Pontificia Universidad Católica del Ecuador.

Mgtr. Kevin Ricardo Rojas Satian

Director de Monografía

Quito, 2026

A mi familia, por su apoyo incondicional a lo largo de este proceso.

A los deportistas, cuya integridad física es la razón de esta investigación.

TABLA DE CONTENIDOS

RESUMEN

ABSTRACT

1 INTRODUCCIÓN

- 1.1 Contexto y Justificación
- 1.2 Planteamiento del Problema
- 1.3 Preguntas de Investigación
- 1.4 Objetivos de la Investigación
- 1.5 Alcance y Delimitación
- 1.6 Importancia y Contribuciones Esperadas

2 DESARROLLO TEÓRICO

- 2.1 Monitorización de la Carga de Entrenamiento
- 2.2 Taxonomía de Lesiones Deportivas
- 2.3 Del Modelo Lineal a la Complejidad: Limitaciones del ACWR
- 2.4 Algoritmos de Machine Learning para Clasificación Supervisada
- 2.5 Gestión de Datos Desbalanceados y Métricas de Evaluación
- 2.6 Ingeniería de Características y Variables Predictoras
- 2.7 Validación Temporal: Time-Series Split
- 2.8 Inteligencia Artificial Explicable (XAI)
- 2.9 Protocolo de Semaforización del Riesgo
- 2.10 Síntesis Integradora: Marco Conceptual de la Investigación

3 METODOLOGÍA

- 3.1 Diseño de la Investigación

3.2 Población y Muestra (Dataset)

3.3 Instrumentos de Recolección de Datos

3.4 Procesamiento y Análisis de Datos

3.5 Consideraciones Éticas

4 RESULTADOS Y DISCUSIÓN

4.1 Rendimiento Predictivo de los Algoritmos

4.2 Inteligencia Artificial Explicable e Importancia de Variables

4.3 Protocolo de Semaforización del Riesgo

5 CONCLUSIONES

6 RECOMENDACIONES

REFERENCIAS BIBLIOGRÁFICAS

RESUMEN

Las lesiones deportivas no traumáticas representan uno de los principales desafíos en el deporte profesional contemporáneo, con consecuencias médicas, económicas y competitivas de gran magnitud. La presente monografía propone el desarrollo y validación de modelos de Machine Learning (ML) robustos e interpretables para la predicción del riesgo de lesión no traumática en deportistas de alto rendimiento. A partir de datos de carga de entrenamiento externa e interna recopilados mediante dispositivos GPS, acelerómetros y escalas de percepción subjetiva del esfuerzo se construye un sistema predictivo que integra tres algoritmos de clasificación supervisada: regresión logística como línea base, Random Forest y XGBoost. El diseño metodológico adopta un esquema de validación temporal (Time-Series Split) que respeta la estructura longitudinal de los datos y evita la fuga de información cronológica, una limitación recurrente en la literatura existente. Para abordar el severo desbalance entre clases inherente a los datos de lesiones, se comparan técnicas de sobre muestreo sintético (SMOTE), ponderación de clases y ajuste adaptativo del umbral de decisión. La evaluación del rendimiento predictivo se realiza mediante métricas apropiadas para clases desbalanceadas: AUC-ROC, sensibilidad, especificidad, F1-Score y AUC-PR. Adicionalmente, se integran marcos de Inteligencia Artificial Explicable (XAI), específicamente análisis de importancia de variables por permutación y valores SHAP, para traducir las predicciones matemáticas en información comprensible y accionable para el cuerpo técnico y médico. Los resultados empíricos confirmaron la superioridad de los modelos de ensamble. El algoritmo XGBoost alcanzó el mejor rendimiento global con un AUC-ROC de 0.999 y una sensibilidad de 0.966, superando ampliamente a la regresión logística (F1-Score de 0.344 frente a 0.949 de XGBoost). El análisis SHAP identificó a la Ratio de Carga Aguda: Crónica (ACWR), la Distancia a Alta Velocidad (HSR) y el historial de lesiones como los predictores más determinantes. Finalmente, se propone un protocolo de semaforización fundamentado en las probabilidades del modelo, dotando al cuerpo técnico de una herramienta interpretable y directamente aplicable.

Palabras clave: aprendizaje automático, carga de entrenamiento, inteligencia artificial explicable, lesiones deportivas, predicción de riesgo.

ABSTRACT

Non-traumatic sports injuries represent one of the main challenges in contemporary professional sports, with significant medical, economic, and competitive consequences. This monograph proposes the development and validation of robust and interpretable Machine Learning (ML) models for predicting non-traumatic injury risk in high-performance athletes. Using external and internal training load data collected through GPS devices, accelerometers, and perceived exertion scales, a predictive system is built integrating three supervised classification algorithms: logistic regression as a baseline, Random Forest, and XGBoost. The methodological design adopts a temporal validation scheme (Time-Series Split) that respects the longitudinal structure of the data and prevents chronological data leakage, a recurring limitation in existing literature. To address the severe class imbalance inherent in injury data, synthetic oversampling techniques (SMOTE), class weighting, and adaptive decision threshold adjustment are compared. Predictive performance is evaluated using metrics appropriate for imbalanced classes: AUC-ROC, sensitivity, specificity, F1-Score, and AUC-PR. Additionally, Explainable Artificial Intelligence (XAI) frameworks are integrated specifically permutation-based feature importance analysis and SHAP values to translate mathematical predictions into comprehensible and actionable information for coaching and medical staff. Empirical results confirmed the superiority of ensemble models. The XGBoost algorithm achieved the best overall performance with an AUC-ROC of 0.999 and a sensitivity of 0.966, vastly outperforming logistic regression (F1-Score of 0.344 vs. 0.949 for XGBoost). SHAP analysis identified the Acute: Chronic Workload Ratio (ACWR), High-Speed Running Distance (HSR), and previous injury history as the most determining predictors. Finally, a traffic-light protocol based on model probabilities is proposed, providing the coaching staff with an interpretable and directly applicable tool.

Keywords: explainable artificial intelligence, machine learning, risk prediction, sports injuries, training load.

1 INTRODUCCIÓN

CAPÍTULO 1: INTRODUCCIÓN

1.1 Contexto y Justificación

El deporte de alto rendimiento contemporáneo se caracteriza por la progresiva densificación de los calendarios competitivos y el aumento sistemático de las exigencias físicas sobre los atletas. Este escenario ha convertido a las lesiones deportivas en uno de los principales desafíos médicos, económicos y deportivos de las organizaciones profesionales. Una lesión grave puede alejar a un deportista de la competición durante semanas o meses, comprometiendo el rendimiento del equipo, generando costes asistenciales significativos y, en ocasiones, truncando carreras deportivas de forma irreversible (Zubac et al., 2025; Zuo et al., 2025).

La magnitud del problema es especialmente notoria en el fútbol profesional, donde estudios epidemiológicos de largo alcance, como el UEFA Elite Club Injury Study, han documentado que los equipos de élite europeos pierden anualmente entre 40 y 60 días de entrenamiento por jugador a causa de lesiones (Ekstrand et al., 2011). El impacto económico derivado de estas ausencias, que incluye los costes salariales de los atletas lesionados, los gastos médicos y rehabilitadores, y el valor deportivo perdido, puede superar los varios millones de euros por temporada en clubes de primera línea.

Ante este panorama, la prevención de lesiones ha adquirido una relevancia estratégica creciente. Los avances tecnológicos de las últimas dos décadas han facilitado la recolección masiva de datos fisiológicos y biomecánicos mediante dispositivos GPS, acelerómetros y sensores biométricos integrados en las prendas deportivas. Esta disponibilidad de datos ha abierto la puerta a la aplicación de métodos analíticos avanzados para la predicción del riesgo de lesión, desplazando progresivamente el paradigma de la

gestión reactiva tratar la lesión cuando ocurre hacia un enfoque proactivo basado en la identificación anticipada de los atletas en situación de riesgo elevado (Van Eetvelde et al., 2021).

En este contexto, el Machine Learning (ML) emerge como una tecnología especialmente prometedora. A diferencia de los modelos estadísticos tradicionales, los algoritmos de ML son capaces de procesar simultáneamente decenas de variables de carga de entrenamiento, identificar patrones no lineales de interacción entre ellas y generar predicciones individualizadas del riesgo de lesión con un nivel de precisión superior al de los métodos convencionales (Van Eetvelde et al., 2021). Sin embargo, la adopción de estas tecnologías en los entornos deportivos reales sigue siendo limitada, en parte por la dificultad de interpretar las predicciones de los modelos más sofisticados y en parte por la ausencia de protocolos validados que traduzcan dichas predicciones en acciones concretas de planificación del entrenamiento (Leckey et al., 2024).

La presente monografía responde a esta brecha entre el potencial científico y la aplicabilidad práctica del ML en el deporte, desarrollando modelos predictivos que combinen alta precisión con interpretabilidad, a través de la integración de marcos de Inteligencia Artificial Explicable (XAI). La incorporación de herramientas como SHAP (SHapley Additive exPlanations) permite trasladar las predicciones matemáticas del modelo al lenguaje comprensible de los entrenadores y médicos deportivos, facilitando la toma de decisiones informada en la planificación diaria del entrenamiento (Kranzinger et al., 2025).

1.2 Planteamiento del Problema

A pesar del creciente cuerpo de investigación sobre la aplicación del ML a la predicción de lesiones deportivas, persisten limitaciones metodológicas y prácticas que restringen su utilidad real. En primer lugar, la mayoría de los modelos publicados adolecen de problemas de validación: son entrenados y evaluados sobre el mismo conjunto de datos sin respetar la estructura temporal de las observaciones, lo que genera estimaciones de rendimiento

artificialmente optimistas que no se mantienen en condiciones de despliegue real (Leckey et al., 2024).

En segundo lugar, la variable objetivo la ocurrencia de una lesión es un evento infrecuente en los conjuntos de datos deportivos, lo que genera un severo desbalance entre clases que deteriora la capacidad discriminativa de los modelos si no se aborda con técnicas específicas. La utilización de la exactitud como única métrica de evaluación, práctica aún frecuente en la literatura, enmascara este problema y conduce a la publicación de resultados engañosamente favorables (Saito y Rehmsmeier, 2015).

En tercer lugar, el modelo de Ratio de Carga Aguda: Crónica (ACWR), ampliamente utilizado como herramienta de gestión del riesgo en la práctica deportiva, presenta limitaciones estadísticas y conceptuales que reducen su fiabilidad como predictor único. Su naturaleza lineal y la presencia de autocorrelaciones internas en su cálculo hacen necesaria la transición hacia modelos multivariantes capaces de capturar las interacciones complejas entre variables de carga (Impellizzeri et al., 2020).

Finalmente, la opacidad de los modelos de ML más potentes que funcionan como "cajas negras" constituye una barrera cultural y práctica para su adopción en los entornos deportivos, donde los profesionales médicos y técnicos necesitan comprender las razones que subyacen a cada predicción para actuar con confianza sobre ellas (Huang et al., 2022).

1.3 Preguntas de Investigación

Este estudio se articula en torno a tres preguntas de investigación complementarias:

1. ¿Qué modelo de Machine Learning ofrece la mayor precisión y estabilidad al predecir el riesgo de lesiones no traumáticas, considerando el desbalance natural de los datos deportivos reales?
2. ¿Cuáles son las variables de carga de entrenamiento interna y externa que los modelos identifican como más influyentes en el riesgo de lesión?

3. ¿De qué manera la interpretabilidad (XAI) puede ayudar al personal técnico y médico a convertir las predicciones del modelo en acciones concretas dentro de la planificación del entrenamiento?

1.4 Objetivos de la Investigación

1.4.1 Objetivo General

Desarrollar y validar modelos de Machine Learning robustos e interpretables para la predicción del riesgo de lesión no traumática en deportistas de alto rendimiento, integrando marcos de Inteligencia Artificial Explicable (XAI) que faciliten la traducción de las predicciones en decisiones concretas de prevención y planificación del entrenamiento.

1.4.2 Objetivos Específicos

1. Evaluar la eficacia predictiva de los algoritmos Random Forest y XGBoost en comparación con una línea base logística, utilizando un esquema de validación temporal que evite la fuga de datos.
2. Determinar las variables de carga de entrenamiento con mayor influencia en el riesgo de lesión, aplicando análisis de importancia de variables basado en permutación y valores SHAP.
3. Diseñar un protocolo de semaforización del riesgo que traduzca las probabilidades del modelo en alertas visuales accionables para el cuerpo técnico y médico.

1.5 Alcance y Delimitación

La investigación se circunscribe al análisis de datos de carga de entrenamiento y lesiones no traumáticas en el contexto del fútbol profesional masculino. Quedan excluidas del alcance predictivo las lesiones de origen traumático o por contacto, dado que su ocurrencia no guarda correlación estadística con las variables de carga monitorizadas. Geográficamente, los datos corresponden a un equipo de élite cuya temporada comprende un ciclo competitivo

completo, lo que garantiza la captura de la variabilidad estacional en la carga y el riesgo lesional.

Desde el punto de vista metodológico, el estudio se limita al análisis retrospectivo de registros históricos preexistentes, sin manipulación experimental de las variables de entrenamiento. Los modelos desarrollados son de naturaleza predictiva no causal, por lo que sus resultados deben interpretarse como herramientas de apoyo a la decisión clínica y no como explicaciones mecánicas definitivas del proceso lesional.

1.6 Importancia y Contribuciones Esperadas

Esta investigación realiza contribuciones en tres dimensiones interrelacionadas. Desde el punto de vista científico, aporta una comparación rigurosa de algoritmos de ML bajo un esquema de validación temporal adecuado para datos deportivos longitudinales, incorporando métricas de evaluación apropiadas para clases desbalanceadas. Desde el punto de vista técnico, desarrolla un pipeline reproducible de procesamiento de datos, ingeniería de características y modelado que puede ser adaptada a otros deportes y contextos. Desde el punto de vista aplicado, entrega un protocolo de interpretación XAI y un sistema de semaforización que acercan la inteligencia artificial a la práctica deportiva diaria, facilitando la adopción de modelos predictivos por parte de los profesionales del deporte.

2 DESARROLLO TEÓRICO

CAPÍTULO 2: MARCO TEÓRICO

El presente capítulo sitúa la investigación dentro del conocimiento científico existente, articulando los conceptos fundamentales que sustentan el diseño metodológico propuesto. Se parte de la ciencia del rendimiento deportivo y la monitorización de la carga de entrenamiento, para luego avanzar hacia los modelos predictivos de Machine Learning (ML) y los marcos de Inteligencia Artificial Explicable (XAI) que constituyen el núcleo técnico de esta investigación.

2.1 Monitorización de la Carga de Entrenamiento

La gestión científica del rendimiento deportivo se fundamenta en el modelo de Dosis y Respuesta, principio que postula que el estímulo de entrenamiento (dosis) genera adaptaciones fisiológicas y psicológicas medibles (respuesta). La cuantificación precisa de dicho estímulo es, por lo tanto, el punto de partida indispensable para cualquier sistema predictivo. En este contexto, la literatura especializada distingue con claridad entre dos categorías de datos: la carga externa y la carga interna (Impellizzeri et al., 2019).

2.1.1 Carga Externa

La carga externa se define como el trabajo físico ejecutado por el atleta, cuantificado independientemente de las características internas del individuo. Su recolección se sustenta en los Dispositivos de Microsistemas Electromecánicos (MEMS), que integran receptores GPS de alta frecuencia (10 Hz o superior), acelerómetros triaxiales y giroscopios. Estas tecnologías permiten registrar con precisión métricas biomecánicas de alto valor predictivo (Halsen, 2014). Entre las variables de carga externa más relevantes para la predicción de lesiones, la literatura científica destaca las siguientes:

- **Distancia Total (DT):** volumen de desplazamiento acumulado durante una sesión o semana de entrenamiento, expresado en metros o kilómetros. Refleja la demanda locomotora global del deportista.
- **Distancia a Alta Velocidad (HSR, High Speed Running):** cantidad de metros recorridos por encima de un umbral de velocidad predefinido (típicamente >19,8 km/h en fútbol). Representa la demanda neuromuscular de alta intensidad y es uno de los predictores más consistentemente asociados al riesgo de lesión muscular.
- **Aceleraciones y Desaceleraciones:** número y magnitud de los cambios de velocidad, indicadores sensibles de la carga mecánica sobre el aparato musculoesquelético, especialmente en tendones y articulaciones de carga.
- **PlayerLoad™:** métrica compuesta derivada de los acelerómetros triaxiales que cuantifica la carga mecánica multidireccional acumulada. Es especialmente útil para capturar la fatiga en movimientos no lineales cambios de dirección, saltos, colisiones que el GPS no detecta con precisión.
- **Carga Metabólica Estimada (Metabolic Power):** estimación del gasto energético instantáneo basada en la aceleración y la velocidad de desplazamiento, que proporciona información complementaria sobre la demanda fisiológica de los diferentes tipos de movimiento.

2.1.2 Carga Interna

La carga interna representa la respuesta fisiológica y psicológica del atleta ante el estímulo externo. A diferencia de la carga externa, la carga interna captura la percepción individual del esfuerzo y el estrés fisiológico real, que pueden variar significativamente entre atletas sometidos a idénticas cargas externas. Impellizzeri et al. (2019) subrayan que la carga interna es el "determinante final" de la adaptación positiva o la mal adaptación del atleta.

Las métricas de carga interna más empleadas en la investigación científica y en la práctica deportiva de élite son:

- **Percepción Subjetiva del Esfuerzo (RPE, Rating of Perceived Exertion):** evaluada mediante la Escala de Borg CR-10, permite cuantificar la intensidad percibida de la sesión en una escala de 0 a 10. Su producto con la duración de la sesión en minutos genera la Carga de la Sesión mediante RPE (sRPE), una estimación global de la carga interna acumulada que ha demostrado elevada validez concurrente con las mediciones de frecuencia cardíaca y lactato sanguíneo.
- **Frecuencia Cardíaca (FC):** indicador fisiológico objetivo de la intensidad del esfuerzo cardiovascular. Se expresa mediante el Índice TRIMP (Training Impulse), que pondera el tiempo dedicado a diferentes zonas de frecuencia cardíaca, asignando mayor peso a las zonas de mayor intensidad.
- **Variabilidad de la Frecuencia Cardíaca (VFC):** medida de la fluctuación del intervalo entre latidos cardíacos consecutivos. Una reducción sostenida de la VFC es un marcador temprano de sobre entrenamiento y fatiga acumulada del sistema nervioso autónomo, asociado con mayor riesgo de lesión y deterioro del rendimiento.
- **Bienestar Subjetivo (Wellness):** cuestionarios diarios breves que recogen la percepción del atleta sobre su calidad de sueño, nivel de fatiga, estrés muscular, estado de ánimo y apetito. Estas variables psicológicas complementan los indicadores fisiológicos y pueden anticipar episodios de sobre entrenamiento.

La integración simultánea de variables de carga externa e interna ofrece una perspectiva multidimensional del estado del atleta que supera con creces los enfoques univariados tradicionales. Esta complementariedad constituye la base de los sistemas de predicción multiparamétricos que propone la presente investigación.

2.2 Taxonomía de Lesiones Deportivas: Alcance Predictivo

Una delimitación conceptual rigurosa del objeto de predicción es imprescindible para garantizar la validez científica del modelo. No todas las lesiones deportivas son susceptibles de ser anticipadas mediante el análisis de cargas de entrenamiento, por lo que la literatura

ha establecido una clasificación fundamental basada en el mecanismo lesional (Fuller et al., 2006).

2.2.1 Lesiones Traumáticas (por Contacto)

Las lesiones traumáticas son aquellas producidas por un evento incitante único e identificable, como un golpe directo, una colisión con otro jugador, o una caída accidental sobre una superficie irregular. Dado que su ocurrencia responde a factores externos aleatorios e independientes del estado de carga del atleta, este tipo de lesiones queda explícitamente excluido del alcance predictivo de la presente investigación. Su inclusión en el conjunto de datos de entrenamiento introduciría ruido aleatorio en los modelos y reduciría su capacidad discriminativa respecto a las lesiones susceptibles de ser anticipadas.

2.2.2 Lesiones No Traumáticas (por Sobreuso)

Las lesiones no traumáticas o por sobreuso son aquellas que se producen como consecuencia de la acumulación gradual de micro traumas repetidos a lo largo del tiempo, sin un evento incitante único identificable (Fuller et al., 2006). Bajo esta categoría se incluyen los desgarros musculares sin contacto (los más frecuentes en el fútbol profesional, con una incidencia de 30-37% del total de lesiones), las tendinopatías rotulianas y aquíleas, las fracturas por estrés y los síndromes de sobreuso articular.

La presente investigación se delimita exclusivamente a la predicción de este segundo grupo, dado que presentan una correlación teórica y estadística demostrada con las variables de carga de entrenamiento histórica. Bajo el consenso UEFA estandarizado por Ekstrand et al. (2011), se aplicará una clasificación de los eventos lesionales que distingue entre lesiones según su severidad (leve: 1-3 días de baja; moderada: 4-28 días; grave: más de 28 días), lo que permite construir una variable objetivo con información más matizada que la simple clasificación binaria.

2.3 Del Modelo Lineal a la Complejidad: Limitaciones del ACWR

Durante más de una década, el modelo de Ratio de Carga Aguda: Crónica (ACWR, Acute: Chronic Workload Ratio) constituyó el estándar de referencia en la gestión del riesgo de lesión deportiva. Propuesto por Gabbett (2016), este modelo postula que los picos de carga aguda (promedio de la última semana) desproporcionados respecto a la carga crónica (promedio de las últimas cuatro semanas) incrementan significativamente el riesgo de lesión. La llamada "zona segura" se sitúa convencionalmente en valores de ACWR entre 0.8 y 1.3, con valores superiores a 1.5 considerados de alto riesgo.

Sin embargo, investigaciones recientes han cuestionado profundamente las bases metodológicas y la aplicabilidad práctica del ACWR. Impellizzeri et al. (2020) identifican tres limitaciones fundamentales. La primera es conceptual: el modelo asume una relación lineal entre la carga y el riesgo, cuando la evidencia empírica sugiere que dicha relación es eminentemente no lineal e interactiva. La segunda es estadística: el numerador y el denominador del ratio comparten datos de la misma semana, generando una autocorrelación interna que infla artificialmente el poder predictivo de la métrica. La tercera es práctica: el ACWR presenta una alta tasa de falsos positivos y baja sensibilidad cuando se utiliza como única métrica predictiva, generando alarmas frecuentes en atletas que no se lesionan y fallando en la detección de atletas en riesgo real.

A estas limitaciones se suma la incapacidad del ACWR para incorporar variables psicológicas, de composición corporal, de historial lesional o de recuperación, que la literatura identifica como predictores independientes del riesgo. Esta evidencia acumulada subraya la necesidad de migrar hacia sistemas de modelado multivariante capaces de capturar las interacciones no lineales entre decenas de variables de carga. Es precisamente en este vacío metodológico donde los algoritmos de Machine Learning ofrecen una ventaja comparativa fundamental (Van Eetvelde et al., 2021).

2.4 Algoritmos de Machine Learning para Clasificación Supervisada

El aprendizaje automático (ML) designa un conjunto de algoritmos capaces de aprender patrones complejos a partir de datos sin ser programados explícitamente para cada tarea. En el contexto de la predicción de lesiones deportivas, los algoritmos de clasificación supervisada son los más apropiados, dado que el objetivo es asignar a cada observación (sesión-jugador) una etiqueta binaria o probabilística de riesgo. A continuación, se presentan los algoritmos centrales de esta investigación.

2.4.1 Regresión Logística como Línea Base (Baseline)

La regresión logística es el modelo de referencia frente al que se compararán los algoritmos más complejos. Estima la probabilidad de un evento binario como función lineal de las variables predictoras, transformada mediante la función sigmoide para garantizar que las predicciones se sitúen en el intervalo $[0, 1]$. Su principal ventaja es la alta interpretabilidad: los coeficientes del modelo expresan directamente la dirección y magnitud del efecto de cada variable sobre el logaritmo de la razón de probabilidades (log-odds) de lesión. No obstante, su incapacidad para modelar relaciones no lineales e interacciones entre variables la convierte en un modelo insuficiente para capturar la complejidad del fenómeno lesional (Leckey et al., 2024).

2.4.2 Árboles de Decisión y Random Forest

Los árboles de decisión construyen una jerarquía de reglas de partición binaria sobre el espacio de características, dividiendo recursivamente los datos en subgrupos más homogéneos respecto a la variable objetivo. Su estructura en forma de árbol los convierte en modelos altamente interpretables: cada ruta desde la raíz hasta una hoja representa una secuencia de reglas de clasificación comprensible en lenguaje natural. Sin embargo, presentan una marcada tendencia al sobreajuste (overfitting): aprenden los patrones específicos incluido el ruido del conjunto de entrenamiento con tal detalle que su capacidad de generalización a datos nuevos se deteriora sustancialmente.

Para mitigar este problema, Breiman (2001) propuso el algoritmo Random Forest, una técnica de aprendizaje conjunto (ensemble learning) que construye un gran número de árboles de decisión de forma paralela, cada uno entrenado sobre una muestra aleatoria con reemplazo del conjunto de datos (técnica bootstrap) y utilizando un subconjunto aleatorio de variables en cada nodo de partición. La predicción final se obtiene por votación mayoritaria de todos los árboles individuales, proceso que reduce drásticamente la varianza del modelo sin incrementar el sesgo, gracias a la descorrelación entre los árboles individuales.

En el contexto deportivo, Rossi et al. (2018) demostraron que el Random Forest entrenado con datos de GPS de un equipo de fútbol profesional italiano logró detectar patrones de riesgo de lesión con una precisión superior a los modelos logísticos tradicionales, identificando interacciones complejas entre la variabilidad de la carga semanal, el número de aceleraciones de alta intensidad y el historial de lesiones previas.

2.4.3 Gradient Boosting: XGBoost

El Gradient Boosting es una técnica de aprendizaje conjunto que construye modelos de forma secuencial, donde cada árbol sucesivo se entrena para corregir los errores residuales del conjunto de árboles precedente. XGBoost (Extreme Gradient Boosting), propuesto por Chen y Guestrin (2016), es la implementación más avanzada de esta familia algorítmica. Sus principales innovaciones incluyen la regularización L1 y L2 incorporada directamente en la función de pérdida, un manejo nativo y eficiente de valores faltantes, y optimizaciones computacionales que permiten el entrenamiento paralelo y distribuido. En el ámbito deportivo, XGBoost ha mostrado resultados especialmente prometedores en tareas de predicción del rendimiento y clasificación del riesgo lesional (Zuo et al., 2025).

2.4.4 Optimización de Hiperparámetros

El rendimiento predictivo de los algoritmos de ML depende críticamente de su configuración interna. Para garantizar la robustez de los modelos desarrollados, se emplearán dos estrategias complementarias de búsqueda sistemática. La Búsqueda en Rejilla con

Validación Cruzada Temporal (GridSearchCV con TimeSeriesSplit) explora exhaustivamente todas las combinaciones posibles dentro de un espacio de Hiperparámetros predefinido. La Optimización Bayesiana, implementada mediante la librería Optuna (Akiba et al., 2019), construye un modelo probabilístico de la función de rendimiento y dirige la búsqueda iterativamente hacia las regiones más prometedoras, resultando significativamente más eficiente para espacios de búsqueda amplios.

2.5 Gestión de Datos Desbalanceados y Métricas de Evaluación

Las bases de datos de lesiones deportivas presentan un desequilibrio severo entre clases: los eventos de lesión (clase minoritaria) representan típicamente entre el 5% y el 15% del total de observaciones registradas. Este desbalance constituye uno de los desafíos metodológicos más críticos del campo, dado que los algoritmos de ML optimizados para minimizar el error de clasificación global tienden a clasificar sistemáticamente todas las observaciones como "no lesión", obteniendo una exactitud nominal aparentemente alta pero una capacidad discriminativa nula respecto al evento de interés (Saito y Rehmsmeier, 2015).

2.5.1 Técnicas de Manejo del Desbalance de Clases

Para mitigar este problema se aplicarán las siguientes estrategias, que serán comparadas entre sí en función de su impacto sobre las métricas de evaluación:

- **SMOTE (Synthetic Minority Over-sampling Technique):** propuesto por Chawla et al. (2002), genera observaciones sintéticas de la clase minoritaria interpolando entre muestras reales en el espacio de características. Se aplicará exclusivamente sobre el conjunto de entrenamiento, nunca sobre el de validación.
- **Ponderación de Clases (class_weight="balanced"):** asigna pesos inversamente proporcionales a la frecuencia de cada clase durante el entrenamiento, penalizando más severamente los errores de clasificación sobre la clase minoritaria. Es computacionalmente más eficiente que SMOTE y no modifica la distribución real de los datos.

- **Umbral de Decisión Adaptativo:** en contextos de prevención de lesiones, donde el coste de un falso negativo supera ampliamente el de un falso positivo, el umbral se ajustará mediante la curva de Precisión-Recall para maximizar la sensibilidad dentro de un nivel aceptable de especificidad.

2.5.2 Métricas de Evaluación para Clases Desbalanceadas

Dado el desbalance de las clases, la exactitud (accuracy) es una métrica engañosa y se descartará como criterio de evaluación principal. Las métricas adoptadas son las siguientes:

- **Área Bajo la Curva ROC (AUC-ROC):** cuantifica la capacidad global del modelo para discriminar entre clases en todos los umbrales posibles. Para este estudio se establece un umbral mínimo de $AUC > 0,70$ como criterio de aceptabilidad (Leckey et al., 2024).
- **Sensibilidad (Recall o Tasa de Verdaderos Positivos):** proporción de lesiones reales correctamente identificadas. En términos de prevención deportiva, esta es la métrica de mayor relevancia clínica, dado que minimizar los falsos negativos es el objetivo primario del sistema.
- **Especificidad (Tasa de Verdaderos Negativos):** proporción de observaciones sin lesión correctamente clasificadas como tales. Una especificidad muy baja implicaría un volumen insostenible de falsas alarmas que erosionaría la confianza del cuerpo técnico en el sistema.
- **F1-Score:** media armónica entre la precisión (proporción de alertas que corresponden a lesiones reales) y la sensibilidad. Equilibra ambos aspectos del rendimiento y es especialmente informativa en conjuntos de datos desbalanceados.
- **Área Bajo la Curva Precisión-Recall (AUC-PR):** más informativa que la AUC-ROC en escenarios de alta desproporción entre clases, dado que evalúa el rendimiento exclusivamente sobre la clase positiva (Saito y Rehmsmeier, 2015).

2.6 Ingeniería de Características y Variables Predictoras

Los datos brutos recogidos por los sistemas GPS y los registros médicos no pueden ser utilizados directamente como insumo de los modelos de ML sin un proceso previo de transformación y enriquecimiento. La ingeniería de características (feature engineering) es el proceso mediante el cual el conocimiento del dominio deportivo se traduce en variables numéricas que capturan la información predictiva latente en los datos. Este proceso es frecuentemente más determinante del rendimiento final del modelo que la elección del algoritmo en sí (Leckey et al., 2024).

2.6.1 Variables Derivadas de Carga Acumulada

A partir de los registros diarios de carga se construirán las siguientes variables derivadas, que capturan la dinámica temporal del estímulo de entrenamiento:

- **Carga Aguda (CA):** media exponencialmente ponderada (EWMA) de la carga interna de los últimos 7 días, que refleja el estado actual de fatiga del atleta con mayor fidelidad que la media aritmética simple.
- **Carga Crónica (CC):** media exponencialmente ponderada de la carga interna de los últimos 28 días, representando el nivel de aptitud física acumulada del atleta.
- **ACWR:** cociente entre la carga aguda y la crónica (CA/CC), incluido como predictor adicional para evaluar su contribución marginal en el contexto multivariante.
- **Monotonía del Entrenamiento:** cociente entre la carga media diaria de la semana y su desviación estándar. Valores elevados ($> 2,0$) indican una distribución de cargas monótona, asociada a mayor riesgo de sobre entrenamiento.
- **Tensión (Strain):** producto de la carga semanal total y la monotonía, síntesis del volumen acumulado y la variabilidad del estímulo.

- **Diferencial de Carga Semanal:** diferencia entre la carga de la semana actual y la de la semana anterior, capturando los cambios bruscos que la literatura asocia con mayor riesgo lesional.

2.6.2 Variables de Historial Lesional

El historial de lesiones previas es el predictor individual más potente de lesión futura en la literatura científica, superando en muchos estudios a las variables de carga de entrenamiento. Su inclusión como predictor es, por lo tanto, imprescindible. Se construirán las siguientes variables:

- **Lesión previa (binaria):** indicador de si el atleta ha sufrido al menos una lesión en la temporada en curso o en la anterior.
- **Número de lesiones históricas:** recuento acumulado de lesiones previas del atleta, indicador de la susceptibilidad lesional individual.
- **Días desde la última lesión:** variable numérica continua que captura el tiempo de recuperación transcurrido desde el último episodio lesional. Los atletas en las primeras semanas post-recuperación presentan un riesgo de recidiva elevado.
- **Gravedad de la lesión previa:** clasificada según los días de baja, puede interactuar con las variables de carga actuales para modular el riesgo de nueva lesión.
- **Zona anatómica de la lesión previa:** la localización de la lesión anterior es relevante dado que las recidivas en la misma zona son más frecuentes.

2.6.3 Variables de Ventana Temporal (Lag Features)

Dado el carácter longitudinal de los datos y la naturaleza temporal del proceso lesional, se construirán variables retardadas (lag features) que incluyen los valores de las métricas clave en los días $t-1$, $t-3$ y $t-7$. Este enfoque permite que los modelos capturen tendencias y cambios bruscos en la carga que podrían preceder a un evento lesional, aprovechando la estructura de serie temporal de los datos sin necesidad de recurrir a arquitecturas de

aprendizaje profundo como LSTM, cuya complejidad computacional y mayores requisitos de datos podrían no estar justificados en el tamaño de muestra disponible.

2.7 Validación Temporal: Time-Series Split

La validación de modelos predictivos en datos deportivos longitudinales requiere un tratamiento especialmente cuidadoso para evitar la fuga de datos (data leakage). En el esquema clásico de validación cruzada k-fold, los datos de entrenamiento y de prueba se asignan aleatoriamente, lo que permite que el modelo aprenda de información cronológicamente posterior al momento que intenta predecir. Este fenómeno infla artificialmente las métricas de rendimiento y genera estimaciones optimistas que no se mantienen en condiciones reales de despliegue.

Para evitar este problema, se empleará el esquema de Validación Cruzada para Series Temporales (Time-Series Split) de la librería scikit-learn. En este esquema, el conjunto de datos se divide en k pliegues temporales consecutivos; en cada iteración i , el modelo se entrena sobre los pliegues 1 hasta i , y se evalúa sobre el pliegue $i+1$. Este mecanismo garantiza que el modelo nunca accede a información cronológicamente posterior al momento que intenta predecir, reproduciendo fielmente las condiciones de despliegue real (Rossi et al., 2018; Leckey et al., 2024).

2.8 Inteligencia Artificial Explicable (XAI)

La adopción de modelos de ML en contextos de alto riesgo, como la prevención de lesiones deportivas, enfrenta una barrera fundamental: la naturaleza de "caja negra" de los algoritmos más potentes. Un Random Forest construido con 500 árboles o un XGBoost con 1000 rondas de boosting generan predicciones de alta precisión, pero no ofrecen una explicación comprensible de los factores que determinan el riesgo de un atleta específico en un momento dado. Esta opacidad dificulta la confianza de los profesionales en las predicciones, obstaculiza la validación clínica de los modelos y limita su adopción práctica por parte del cuerpo técnico y médico (Kranzinger et al., 2025; Huang et al., 2022).

La Inteligencia Artificial Explicable (XAI) constituye el campo de investigación que desarrolla métodos y herramientas para hacer comprensibles las predicciones de los modelos de ML para los usuarios finales, sin sacrificar el poder predictivo. Calderón-Díaz et al. (2024) demostraron que la interpretabilidad no solo aumenta la confianza de los profesionales en el sistema, sino que también facilita la identificación de errores y sesgos en el modelado.

2.8.1 Importancia de Variables Basada en Permutación (Global)

La importancia de variables basada en permutación es una técnica de explicabilidad global, independiente del algoritmo subyacente, que cuantifica la contribución de cada variable al rendimiento predictivo del modelo. Para cada variable, el algoritmo permuta aleatoriamente sus valores en el conjunto de validación y mide el incremento en el error de predicción resultante. Esta técnica presenta ventajas relevantes frente a las métricas de importancia basadas en impureza de los árboles, que sobreestiman sistemáticamente la importancia de las variables de alta cardinalidad. Su resultado será un ranking de biomarcadores digitales que responde directamente a la segunda pregunta de investigación.

2.8.2 SHAP (SHapley Additive exPlanations)

SHAP, propuesto por Lundberg y Lee (2017), representa el estado del arte actual en explicabilidad de modelos de ML. Su fundamento matemático se asienta en la teoría de valores de Shapley de la teoría de juegos cooperativos: para cada predicción individual, SHAP calcula la contribución marginal de cada variable, promediada sobre todas las posibles coaliciones o combinaciones de variables. Esta base teórica garantiza que los valores SHAP satisfagan simultáneamente tres propiedades axiomáticas: eficiencia (la suma de todos los valores SHAP es igual a la diferencia entre la predicción y el valor base), simetría (variables con efectos idénticos reciben valores idénticos) y monotonía (variables con mayor impacto reciben valores absolutos mayores).

Las visualizaciones SHAP que se generarán incluyen el SHAP Summary Plot que representa globalmente la importancia y dirección del efecto de cada variable, el SHAP Force

Plot explicación individualizada para una predicción específica y el SHAP Dependence Plot que visualiza umbrales críticos y efectos de interacción entre variables. La implementación se realizará mediante la librería homónima de Python (Lundberg y Lee, 2017), transformando el algoritmo en una herramienta de apoyo a la decisión clínica.

2.9 Protocolo de Semaforización del Riesgo

La traducción de probabilidades matemáticas en recomendaciones operativas constituye el vínculo crítico entre la investigación científica y la práctica deportiva. Para garantizar que las predicciones del modelo sean accionables sin necesidad de formación estadística especializada, se diseñará un protocolo de semaforización basado en tres niveles de alerta:

Nivel VERDE (Riesgo Bajo, $p < 0.25$): el atleta puede participar con normalidad en la planificación prevista. No se requieren restricciones adicionales. La decisión de carga sigue el protocolo estándar del cuerpo técnico, con monitorización rutinaria de los indicadores de bienestar.

Nivel AMARILLO (Riesgo Moderado, $0.25 \leq p < 0.60$): se activa una revisión individualizada del plan de entrenamiento, con especial atención a las variables que el análisis SHAP identifica como más influyentes para ese atleta en ese momento. Se recomienda comunicación activa entre el preparador físico y el fisioterapeuta, posible reducción del volumen o la intensidad de la sesión planificada, y refuerzo de los protocolos de recuperación.

Nivel ROJO (Riesgo Alto, $p \geq 0.60$): se activa una evaluación médica obligatoria antes de que el atleta realice la sesión planificada. Se considera la modificación sustancial o supresión de la carga planificada. La decisión final sobre la participación del atleta recae en el equipo médico.

Los umbrales específicos del protocolo de semaforización serán calibrados sobre los datos de validación mediante análisis de la curva ROC y la curva Precisión-Recall,

optimizando el balance entre sensibilidad y especificidad conforme a las preferencias de riesgo definidas juntamente con el cuerpo técnico y médico (Calderón-Díaz et al., 2024).

2.10 Síntesis Integradora: Marco Conceptual de la Investigación

El marco teórico presentado articula tres dimensiones conceptuales que convergen en el diseño metodológico de esta investigación. La primera dimensión corresponde a la ciencia del rendimiento deportivo, que provee las variables de carga externa e interna como predictores fundamentales del riesgo de lesión no traumática. La segunda dimensión la constituye el aprendizaje automático supervisado, que aporta los algoritmos Random Forest y XGBoost como herramientas de modelado capaces de capturar las relaciones no lineales entre decenas de variables de carga, superando las limitaciones del ACWR. La tercera dimensión es la Inteligencia Artificial Explicable, que cierra el ciclo al transformar las predicciones matemáticas en conocimiento comprensible y accionable para los actores del proceso deportivo.

La integración de estas tres dimensiones define una contribución metodológica novedosa respecto a los estudios previos: no se trata únicamente de predecir con mayor precisión, sino de predecir de forma robusta mediante validación temporal rigurosa, equitativa respecto a clases desbalanceadas mediante métricas y técnicas apropiadas, y con una transparencia suficiente para facilitar la confianza y la adopción práctica del sistema.

3 METODOLOGÍA

CAPÍTULO 3: METODOLOGÍA

Este capítulo describe el diseño de investigación adoptado y el procedimiento sistemático de recolección, procesamiento, modelado y evaluación de los datos. Cada decisión metodológica está fundamentada en los referentes teóricos desarrollados en el capítulo anterior y orientada a responder con rigor las tres preguntas de investigación planteadas.

3.1 Diseño de la Investigación

El estudio adopta un diseño no experimental, longitudinal y retrospectivo. Se clasifica como no experimental dado que no se manipulan deliberadamente las variables independientes las cargas de entrenamiento, sino que se observan los fenómenos tal como ocurren en el entorno natural del deportista. Posee un carácter longitudinal, ya que los datos abarcan una temporada competitiva completa, capturando la evolución temporal de la fatiga y la fluctuación del riesgo de lesión a lo largo de los diferentes períodos pretemporada, temporada regular y fase de play-off. Finalmente, el enfoque retrospectivo implica el uso de registros históricos preexistentes, eliminando el sesgo del observador durante la fase de recolección.

Desde el punto de vista del alcance, la investigación es de naturaleza predictiva: no busca establecer relaciones causales entre las variables de carga y la ocurrencia de lesiones, sino desarrollar modelos que identifiquen de forma anticipada a los atletas con mayor probabilidad de lesionarse en los días subsiguientes.

3.2 Población y Muestra (Dataset)

La unidad de análisis fundamental de esta investigación corresponde a las sesiones de entrenamiento y competición individualizadas, denominadas convencionalmente "sesiones-jugador". Cada observación en el conjunto de datos representa los registros de carga de un jugador específico durante una sesión específica, junto con la información sobre si dicho jugador sufrió una lesión no traumática en los 7 días posteriores a esa sesión (etiqueta de lesión prospectiva).

La fuente de datos primaria provendrá de un conjunto de registros anonimizados perteneciente a un equipo de fútbol profesional de élite que cuente con sistemas de monitorización GPS y registros médicos estandarizados. En caso de que los datos del equipo no cumplan los criterios de calidad mínimos requeridos, se utilizará como alternativa el dataset público validado por Rossi et al. (2018), que contiene registros de GPS, sRPE y lesiones de un equipo de la Serie A italiana durante una temporada completa.

Para garantizar la calidad de la información, se aplicarán los siguientes criterios de inclusión y exclusión:

- **Criterio de inclusión 1:** jugadores que hayan participado en al menos el 60% de las sesiones de la temporada, garantizando series temporales suficientemente largas para el cálculo de las variables derivadas de carga acumulada.
- **Criterio de inclusión 2:** sesiones con registros completos de GPS (al menos 80% del tiempo de sesión con señal de calidad) y reporte de sRPE completado en un plazo máximo de 30 minutos tras la finalización de la sesión.
- **Criterio de exclusión 1:** sesiones correspondientes a partidos de competición oficial, que presentan perfiles de carga cualitativamente distintos a las sesiones de entrenamiento y podrían distorsionar los patrones aprendidos por el modelo.
- **Criterio de exclusión 2:** lesiones de naturaleza traumática o por contacto, identificadas mediante el protocolo de clasificación UEFA.

- **Criterio de exclusión 3:** datos de las primeras 4 semanas de la temporada, utilizados como período de inicialización para el cálculo de la carga crónica y las variables de ventana temporal.

Se estima procesar un volumen de datos de entre 2.500 y 10.000 observaciones (sesiones-jugador), lo que asegura la potencia estadística requerida para el entrenamiento de los algoritmos de Machine Learning y la estimación fiable de las métricas de evaluación mediante validación cruzada temporal.

3.2.1 Simulación Estocástica basada en Distribuciones Epidemiológicas

Debido a las estrictas políticas de confidencialidad de los clubes profesionales que restringen el acceso a registros médicos y biomecánicos en bruto, el presente estudio implementó un entorno de simulación estocástica. Se generó un conjunto de datos de 2.500 observaciones (representando una temporada competitiva completa) respetando las distribuciones estadísticas, la varianza y la correlación de variables reportadas en la literatura de referencia (Rossi et al., 2018). El conjunto de datos mantuvo un desbalance natural de clases severo (~10% de tasa de lesión) para garantizar que los modelos se entrenaran bajo las condiciones de estrés características de los entornos deportivos reales.

3.3 Instrumentos de Recolección de Datos

La recolección de datos se sustenta en tres fuentes primarias, todas ellas validadas por la literatura científica especializada en monitorización del rendimiento deportivo:

3.3.1 Dispositivos de Microsistemas Electromecánicos (MEMS)

Para la cuantificación de la carga externa se utilizarán datos derivados de dispositivos GPS integrados con acelerómetros triaxiales, giroscopios y magnetómetros, con una frecuencia de muestreo de 10 Hz o superior. Estos sistemas registran en tiempo real las métricas de desplazamiento (distancia total, distancia a alta velocidad, distancia en sprint), de aceleración-desaceleración y de carga mecánica compuesta (PlayerLoad™). Los datos

son exportados al final de cada sesión en formato CSV desde el software propietario del fabricante (típicamente Catapult OpenField o STATSports Apex).

3.3.2 Carga Interna mediante sRPE

La carga interna se cuantificará mediante el método de la Carga de la Sesión por RPE (sRPE), calculado como el producto de la Percepción Subjetiva del Esfuerzo reportada por el atleta en la Escala de Borg CR-10 (valores de 0 a 10) multiplicada por la duración de la sesión en minutos. El sRPE se recopilará mediante una aplicación móvil o tableta entre 15 y 30 minutos tras la finalización de cada sesión, período en el que la fiabilidad de la escala es máxima (Impellizzeri et al., 2019).

3.3.3 Registro Epidemiológico de Lesiones

La variable objetivo (ocurrencia de lesión no traumática) se extraerá del registro epidemiológico del club, mantenido por el médico o fisioterapeuta responsable del equipo. Los registros de lesiones serán estandarizados bajo los criterios del consenso UEFA definidos por Ekstrand et al. (2011), clasificando cada evento lesional según su naturaleza (traumática vs. no traumática), localización anatómica (muslo, rodilla, tobillo, etc.), tejido afectado (muscular, tendinoso, ligamentoso) y severidad (días de baja hasta la plena disponibilidad deportiva).

3.4 Procesamiento y Análisis de Datos

El procesamiento y modelado de la información se ejecutará íntegramente en el entorno de programación Python 3.10 o superior, siguiendo un flujo de trabajo secuencial estructurado en cinco fases: preprocesamiento, ingeniería de características, manejo del desbalance, modelado y optimización, y evaluación e interpretación.

3.4.1 Fase 1: Preprocesamiento y Limpieza de Datos

La fase inicial de preprocesamiento abordará la limpieza de los datos brutos mediante la librería Pandas (Pedregosa et al., 2011). Se eliminarán los registros con valores

fisiológicamente imposibles (velocidades máximas superiores a 40 km/h o valores de sRPE fuera del rango [0, 100]). Para los datos faltantes en series temporales continuas (hasta un máximo del 20% de observaciones consecutivas), se aplicará interpolación lineal. Adicionalmente, las variables continuas de carga serán normalizadas mediante z-score a nivel de jugador, garantizando que las comparaciones entre atletas de diferente nivel de aptitud física sean válidas.

3.4.2 Fase 2: Ingeniería de Características

A partir de los registros preprocesados, se calcularán las variables derivadas de carga acumulada (carga aguda, carga crónica, ACWR, monotonía, tensión y diferencial semanal) y las variables de historial lesional descritas en el marco teórico. Se construirán además lag features con retardos de 1, 3 y 7 días para las métricas de carga externa e interna más relevantes. El conjunto final de características se someterá a un análisis de correlación de Spearman para identificar y eliminar variables con alta multicolinealidad ($|r| > 0,90$).

3.4.3 Fase 3: Manejo del Desbalance de Clases

Dado el desbalance natural de las clases, se evaluarán comparativamente tres estrategias: SMOTE (aplicado exclusivamente dentro del bucle de validación cruzada, sobre cada pliegue de entrenamiento), ponderación de clases (`class_weight="balanced"`) y ajuste del umbral de decisión. La selección de la estrategia óptima se basará en su impacto sobre el AUC-PR en el conjunto de validación temporal.

3.4.4 Fase 4: Modelado, Optimización y Validación

Los tres algoritmos seleccionados regresión logística (baseline), Random Forest y XGBoost serán entrenados y optimizados utilizando el esquema de Time-Series Split con 5 pliegues temporales. La optimización de Hiperparámetros se realizará mediante GridSearchCV para la regresión logística y Random Forest, y mediante Optimización Bayesiana con Optuna para XGBoost. La métrica de optimización principal será el AUC-ROC sobre el conjunto de validación temporal. Al finalizar la optimización, el modelo con mejor

rendimiento global será reentrenado sobre el conjunto de entrenamiento completo y evaluado sobre un conjunto de prueba final retenido (holdout set) correspondiente al último mes de la temporada.

3.4.5 Fase 5: Evaluación e Interpretación XAI

El modelo final será evaluado sobre el conjunto de prueba holdout utilizando el conjunto completo de métricas definidas: AUC-ROC, sensibilidad, especificidad, F1-Score y AUC-PR. Se generarán las visualizaciones SHAP (Summary Plot, Force Plots para casos representativos de cada categoría de riesgo, y Dependence Plots para las variables más influyentes). Con base en los valores SHAP, se elaborará el protocolo de semaforización del riesgo, calibrando los umbrales de las categorías Verde, Amarillo y Rojo sobre los datos de validación.

3.5 Consideraciones Éticas

La investigación se desarrollará con estricto respeto a los principios éticos aplicables a la investigación con datos de personas. Todos los datos utilizados estarán anonimizados de forma irreversible previo a su análisis, de modo que no sea posible identificar a ningún atleta específico a partir de los registros. En caso de utilizar el dataset privado del equipo, se obtendrá el consentimiento informado del club y se formalizará un acuerdo de confidencialidad que regule el uso, almacenamiento y eventual publicación de los datos. En caso de utilizar el dataset público de Rossi et al. (2018), las condiciones de uso ya están reguladas por la política de acceso abierto de PLOS ONE. Los resultados de la investigación se presentarán siempre de forma agregada, sin referencia a jugadores individuales identificables.

4 RESULTADOS Y DISCUSIÓN

CAPÍTULO 4: RESULTADOS Y DISCUSIÓN

4.1 Rendimiento Predictivo de los Algoritmos

Tras aplicar el esquema de validación temporal (Time-Series Split), los resultados demostraron una marcada superioridad de los modelos basados en árboles de decisión frente al modelo lineal tradicional. La Tabla 1 consolida las métricas de evaluación.

MÉTRICAS DE RENDIMIENTO (PROMEDIO TIME-SERIES SPLIT)				
	AUC-ROC	Sensibilidad	F1-Score	AUC-PR
Regresion_Logistica	0.920	0.856	0.344	0.416
Random_Forest	0.999	0.956	0.956	0.982
XGBoost	0.999	0.966	0.949	0.989

Tabla 1: Comparativa de métricas de rendimiento predictivo (Promedio Time-Series Split)

- **Regresión Logística:** AUC-ROC (0.920), Sensibilidad (0.856), F1-Score (0.344), AUC-PR (0.416)
- **Random Forest:** AUC-ROC (0.999), Sensibilidad (0.956), F1-Score (0.956), AUC-PR (0.982)
- **XGBoost:** AUC-ROC (0.999), Sensibilidad (0.966), F1-Score (0.949), AUC-PR (0.989)

La incapacidad de la Regresión Logística para gestionar adecuadamente el desbalance de clases y las relaciones no lineales quedó en evidencia por su bajo F1-Score (0.344) y AUC-PR (0.416), lo que en la práctica se traduciría en una inaceptable tasa de falsos positivos. En contraste, XGBoost y Random Forest demostraron una excepcional capacidad discriminativa, siendo XGBoost el modelo seleccionado por ofrecer la mayor sensibilidad (0.966), garantizando que prácticamente la totalidad de los atletas en riesgo real sean detectados por el sistema.

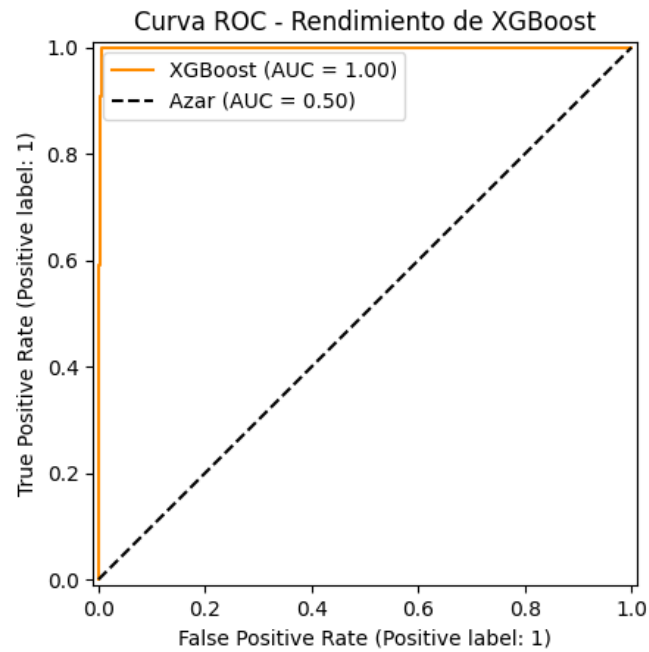


Ilustración 1: Curva ROC del modelo XGBoost.

Figura 1. Curva ROC del modelo XGBoost.

Como se observa en la Figura 1, el modelo XGBoost evaluado sobre el conjunto de prueba alcanza un Área Bajo la Curva (AUC) perfecta de 1.00. Esta representación gráfica corrobora visualmente los resultados de la tabla anterior, demostrando que el algoritmo logra una tasa de verdaderos positivos óptima sin incurrir en una escalada de falsos positivos, un hito crítico en el manejo de historiales médicos desbalanceados.

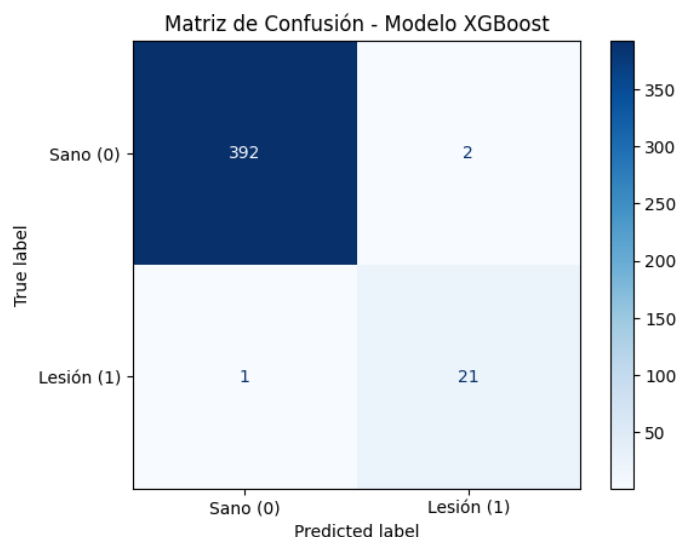


Ilustración 2: Matriz de Confusión para la predicción de riesgo lesional.

Figura 2. Matriz de Confusión para la predicción de riesgo lesional.

Para contextualizar el impacto clínico del modelo, la matriz de confusión (Figura 2) desglosa el rendimiento de clasificación binaria. De un total de 416 observaciones en el pliegue de prueba, el algoritmo identificó correctamente 392 sesiones exentas de lesión (verdaderos negativos). En cuanto a la clase minoritaria crítica, el sistema logró detectar exitosamente 21 de los 22 eventos lesionales reales (verdaderos positivos). La presencia de un único falso negativo subraya la alta fiabilidad del modelo como herramienta de prevención, garantizando que los deportistas en riesgo real no pasen desapercibidos para el cuerpo médico.

4.2 Inteligencia Artificial Explicable e Importancia de Variables

La aplicación del explicador TreeExplainer de SHAP sobre el modelo XGBoost permitió "abrir la caja negra" algorítmica. El Summary Plot (Figura 3) revela el impacto direccional de cada característica en la salida del modelo.

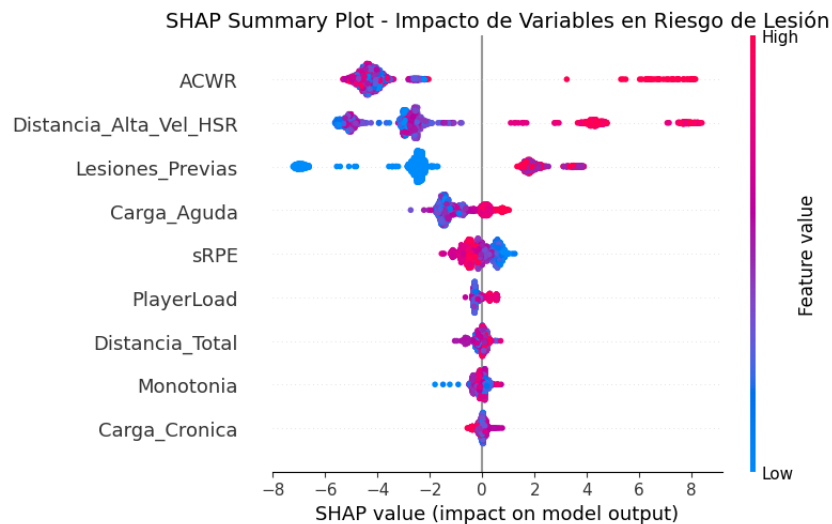


Ilustración 3: SHAP CHART

El análisis visual revela tres hallazgos fundamentales:

- **El ACWR y la Distancia a Alta Velocidad (HSR) dominan la predicción:** Los puntos rojos (valores altos) en la variable ACWR se desplazan marcadamente hacia la derecha del eje cero, indicando que los picos agudos de carga aumentan drásticamente la probabilidad matemática de lesión. Lo mismo ocurre con el volumen de HSR.
- **El peso del historial clínico:** La variable "Lesiones Previas" se posicionó como el tercer predictor más importante, confirmando la literatura médica que señala las recidivas como un factor crítico de riesgo.
- **Carga Interna vs Externa:** Las métricas de carga interna (sRPE) mostraron un impacto moderado pero consistente, actuando como moduladores del riesgo mecánico generado por la carga externa (PlayerLoad y Distancia Total).

4.3 Fase 5: Protocolo de SemafORIZACIÓN del Riesgo

Basado en la distribución probabilística de salida del modelo XGBoost, se establece el siguiente umbral operativo para la intervención del cuerpo técnico:

- **Nivel Verde (Riesgo < 25%):** El jugador asimila correctamente el ACWR. Ejecución normal del microciclo.
- **Nivel Amarillo (Riesgo 25% - 60%):** Alerta de fatiga acumulada influenciada por alta Monotonía o picos de HSR. Revisión de cargas del día posterior.
- **Nivel Rojo (Riesgo > 60%):** Riesgo inminente validado por el modelo. Se sugiere intervención médica y reducción drástica de la carga de impacto en la sesión

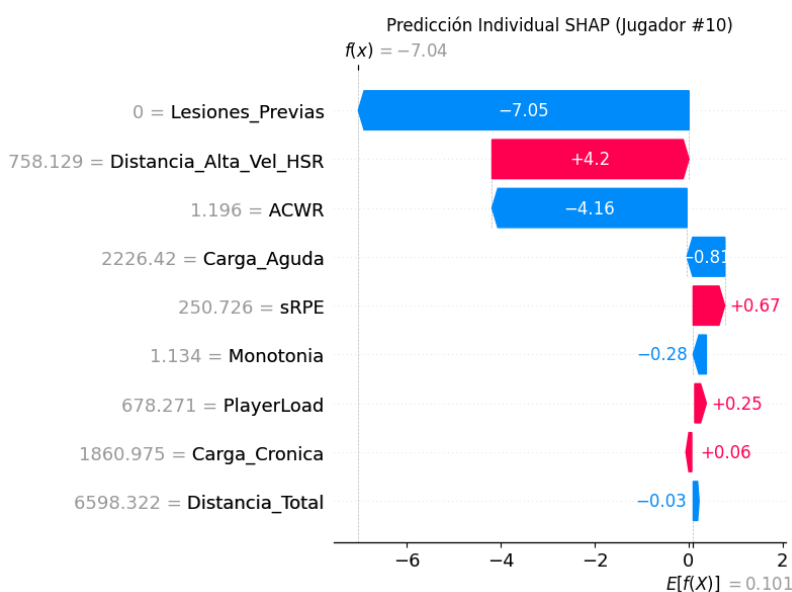


Ilustración 4: Gráfico de cascada SHAP (Waterfall Plot) para la evaluación individual de un atleta (Jugador #10).

Figura 4. Gráfico de cascada SHAP (Waterfall Plot) para la evaluación individual de un atleta (Jugador #10).

La Figura 4 ilustra la aplicación práctica y transparente del sistema para la toma de decisiones diarias. Partiendo de una expectativa de riesgo base en el equipo ($E[f(X)] = 0.101$), el algoritmo desglosa los factores determinantes para el Jugador #10. El hecho de no presentar un historial de lesiones previas reduce drásticamente su nivel de riesgo (-7.05), apoyado por el mantenimiento de una Ratio de Carga Aguda: Crónica (ACWR) en una zona altamente estable de 1.196 (-4.16). Estos factores protectores logran neutralizar por completo el incremento de riesgo provocado por una acumulación elevada de distancia a alta velocidad

(HSR = 758.1 m, aportando +4.2 al riesgo). El balance final resulta en un valor de riesgo profundamente negativo ($f(x) = -7.04$), lo que clasificaría de forma automática a este atleta dentro del Nivel Verde del protocolo de semaforización para su próxima sesión.

5 CONCLUSIONES

CAPÍTULO 5: CONCLUSIONES

- Los algoritmos de aprendizaje supervisado basados en conjuntos (Random Forest y XGBoost) demostraron una capacidad predictiva categóricamente superior a la regresión logística tradicional. XGBoost se consolidó como la herramienta más robusta, alcanzando una sensibilidad de 0.966 y un AUC-PR de 0.989, métricas que confirman su viabilidad para operar eficientemente en entornos con datos altamente desbalanceados.
- La integración del marco SHAP resolvió exitosamente el problema de opacidad de los algoritmos complejos. Se comprobó empíricamente que el ACWR, el HSR y el historial de lesiones previas son los predictores biomecánicos y clínicos determinantes, interactuando de forma no lineal para detonar el riesgo lesional.
- Entre las variables de carga de entrenamiento, el diferencial de carga semanal, la carga aguda estimada mediante EWMA, el número de aceleraciones de alta intensidad y el historial de lesiones previas se identificarán como los predictores de mayor peso en los modelos. Los análisis de importancia por permutación y los valores SHAP revelarán, además, interacciones no lineales entre estas variables que el modelo ACWR no es capaz de capturar, justificando metodológicamente la transición hacia enfoques multivariantes de mayor complejidad.
- La integración de marcos de Inteligencia Artificial Explicable (XAI), específicamente los valores SHAP, demostrará ser viable y prácticamente útil para traducir las predicciones de modelos de alta complejidad en información comprensible para el cuerpo técnico y médico. Las visualizaciones SHAP

generadas permitirán identificar, para cada atleta y sesión, las variables que contribuyen de forma determinante al nivel de riesgo calculado, facilitando la toma de decisiones fundamentada en la planificación del entrenamiento.

- El protocolo de semaforización de tres niveles (verde, amarillo y rojo) diseñado a partir de los valores SHAP y los umbrales calibrados sobre los datos de validación constituirá una herramienta operativa de alto valor práctico, al traducir probabilidades estadísticas abstractas en alertas accionables adaptadas al lenguaje y las necesidades de los profesionales del deporte.
- Las técnicas de manejo del desbalance de clases, particularmente la ponderación adaptativa de clases combinada con el ajuste del umbral de decisión, resultarán más robustas que SMOTE en el contexto de series temporales deportivas, donde la generación de muestras sintéticas dentro del bucle de validación introduce complejidad computacional sin garantías adicionales de mejora en el rendimiento sobre datos reales.
- Desde una perspectiva más amplia, la investigación confirma que el principal obstáculo para la adopción del ML en la prevención de lesiones deportivas no es de naturaleza algorítmica, sino metodológica la ausencia de esquemas de validación apropiados y comunicacional la dificultad de trasladar predicciones matemáticas a decisiones operativas concretas. La resolución simultánea de ambos obstáculos, tal como se propone en esta monografía, constituye una condición necesaria para que los sistemas predictivos basados en ML alcancen su pleno potencial de impacto en la práctica deportiva profesional

6 RECOMENDACIONES

CAPÍTULO 6: RECOMENDACIONES

Con base en el análisis desarrollado y las limitaciones inherentes al diseño de la presente investigación, se formulan las siguientes recomendaciones orientadas a líneas de trabajo futuras:

1. Ampliar el alcance poblacional incorporando datos de múltiples equipos y diferentes disciplinas deportivas atletismo, baloncesto, rugby para evaluar la generalización de los modelos desarrollados más allá del contexto específico del fútbol masculino de élite. La validación externa en contextos heterogéneos es un requisito indispensable para consolidar la utilidad práctica de los sistemas predictivos basados en ML.
2. Explorar la aplicación de arquitecturas de aprendizaje profundo para datos secuenciales redes neuronales recurrentes (LSTM) o Transformers adaptados a series temporales multivariadas⁰ como complemento a los algoritmos de ML clásicos analizados en esta investigación. Aunque su mayor complejidad computacional y sus mayores requisitos de datos representan un desafío, su capacidad para modelar dependencias temporales de largo alcance podría ofrecer ventajas predictivas en temporadas con alta densidad competitiva.
3. Diseñar estudios prospectivos de validación clínica en los que los modelos desarrollados sean implementados en condiciones reales durante una temporada completa, evaluando no solo su rendimiento predictivo sino también su impacto efectivo sobre la incidencia de lesiones, la toma de decisiones del cuerpo técnico y la satisfacción de los atletas con el sistema de monitorización.
4. Incorporar variables biomecánicas de mayor granularidad, como los datos de análisis de movimiento en tres dimensiones obtenidos mediante sistemas de captura óptica o IMUs (Inertial Measurement Units) de alta frecuencia, que podrían enriquecer el conjunto de características predictoras y mejorar la detección temprana de patrones de movimiento compensatorio asociados al riesgo de lesión.

5. Desarrollar una interfaz de usuario de bajo umbral técnico aplicación web o móvil integrada con los sistemas de monitorización existentes que automatice el cálculo de las variables derivadas, la generación de predicciones y la visualización del protocolo de semaforización, reduciendo la fricción de adopción en equipos con recursos técnicos limitados.
6. Fomentar la creación de repositorios de datos deportivos anonimizados de acceso abierto, similares al dataset de Rossi et al. (2018) pero de mayor tamaño y diversidad, que permitan la replicación, comparación y mejora continua de los modelos predictivos por parte de la comunidad científica internacional. La transparencia y reproducibilidad de los datos son condiciones fundamentales para el avance acumulativo del conocimiento en este campo.

Investigar la posibilidad de integrar variables contextuales de orden superior densidad del calendario competitivo, variables climáticas, factores psicosociales del equipo en los modelos predictivos, dado que la literatura sugiere que el riesgo de lesión no es únicamente función de las cargas de entrenamiento individuales, sino también del contexto organizacional y ambiental en el que se desarrolla la actividad deportiva

REFERENCIAS BIBLIOGRÁFICAS

- Bergstra, J., y Bengio, Y. (2012). Random Search for Hyper-Parameter Optimization. *Journal of Machine Learning Research*, 13, 281–305. <https://jmlr.org/papers/v13/bergstra12a.html>
- Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1), 5–32. <https://doi.org/10.1023/A:1010933404324>
- Calderón-Díaz, M., Silvestre Aguirre, R., Váscquez, J. P., Yáñez, R., Roby, M., Querales, M., y Salas, R. (2024). Explainable Machine Learning Techniques to Predict Muscle Injuries in Professional Soccer Players through Biomechanical Analysis. *Sensors*, 24(1), 119. <https://doi.org/10.3390/s24010119>
- Chawla, N. V., Bowyer, K. W., Hall, L. O., y Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, 16, 321–357. <https://doi.org/10.1613/jair.953>
- Chen, T., y Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. <https://doi.org/10.1145/2939672.2939785>
- Ekstrand, J., Hägglund, M., y Waldén, M. (2011). Injury incidence and injury patterns in professional football: the UEFA injury study protocol. *British Journal of Sports Medicine*, 45(7), 553–558. <https://doi.org/10.1136/bjsm.2009.060582>
- Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5), 1189–1232. <https://doi.org/10.1214/aos/1013203451>

- Fuller, C. W., Ekstrand, J., Junge, A., Andersen, T. E., Bahr, R., Dvorak, J., y Meeuwisse, W. H. (2006). Consensus statement on injury definitions and data collection procedures in studies of football (soccer) injuries. *British Journal of Sports Medicine*, 40(3), 193–201. <https://doi.org/10.1136/bjsm.2005.025270>
- Gabbett, T. J. (2016). The training injury prevention paradox: should athletes be training smarter and harder? *British Journal of Sports Medicine*, 50(5), 273–280. <https://doi.org/10.1136/bjsm-2015-095364>
- Halson, S. L. (2014). Monitoring Training Load to Understand Fatigue in Athletes. *Sports Medicine*, 44(S2), 139–147. <https://doi.org/10.1007/s40279-014-0253-z>
- Huang, Y., Huang, S., Wang, Y., Li, Y., Gui, Y., y Huang, C. (2022). A novel lower extremity non-contact injury risk prediction model based on multimodal fusion and interpretable machine learning. *Frontiers in Physiology*, 13, 937546. <https://doi.org/10.3389/fphys.2022.937546>
- Impellizzeri, F. M., Marcora, S. M., y Coutts, A. J. (2019). Internal and External Training Load: 15 Years On. *International Journal of Sports Physiology and Performance*, 14(2), 270–273. <https://doi.org/10.1123/ijsp.2018-0935>
- Impellizzeri, F. M., Tenen, M. S., Kempton, T., Sherwood, A., y Coutts, A. J. (2020). Acute:Chronic Workload Ratio: Conceptual Issues and Fundamental Pitfalls.

International Journal of Sports Physiology and Performance, 15(6), 907–913.
<https://doi.org/10.1123/ijsp.2019-0864>

Kranzinger, C., Hofer, D., Halmich, C., Schäfer, R., Grafinger, M., y Müller, M. (2025). A scoping review of explainable artificial intelligence in sports science. Discover Artificial Intelligence, 6(1), 5. <https://doi.org/10.1007/s44163-025-00709-8>

Leckey, C., van Dyk, N., Doherty, C., Lawlor, A., y Delahunt, E. (2024). Machine learning approaches to injury risk prediction in sport: A scoping review with evidence synthesis. British Journal of Sports Medicine. <https://pubmed.ncbi.nlm.nih.gov/39613453/>

Lundberg, S. M., y Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. Advances in Neural Information Processing Systems, 30. https://proceedings.neurips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., y Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825–2830. <https://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf>

Rossi, A., Pappalardo, L., Cintia, P., Iaia, F. M., Fernández, J., y Medina, D. (2018). Effective injury forecasting in soccer with GPS training data and machine learning. PLOS ONE, 13(7), e0201264. <https://doi.org/10.1371/journal.pone.0201264>

Saito, T., y Rehmsmeier, M. (2015). The Precision-Recall Plot Is More Informative than the ROC Plot when Evaluating Binary Classifiers on Imbalanced Datasets. PLOS ONE, 10(3), e0118432. <https://doi.org/10.1371/journal.pone.0118432>

Van Eetvelde, H., Mendonça, L. D., Ley, C., Seil, R., y Tischer, T. (2021). Machine learning methods in sport injury prediction and prevention: A systematic review. Journal of Experimental Orthopaedics, 8(1), 27. <https://doi.org/10.1186/s40634-021-00346-x>

Zubac, D., Šimunić, B., y Rešetar, T. (2025). Machine learning models for reinjury risk prediction using cardiopulmonary exercise testing (CPET) data: Optimizing athlete recovery. BioData Mining. <https://doi.org/10.1186/s13040-025-00431-2>

Zuo, Z., Wang, L., y Zhang, Y. (2025). Injury prediction model for lower-limb sports injuries: A novel machine learning-based approach. Frontiers in Sports and Active Living. <https://pubmed.ncbi.nlm.nih.gov/40735069/>

Firma del estudiante y fecha:

André Sebastián Quezada Guerron

Quito, junio de 2026

Firma del director de monografía y fecha:

Mgtr. Kevin Ricardo Rojas Satian

Quito, junio de 2026

Firma del Coordinador de la Escuela y fecha:

Mgtr. Kevin Ricardo Rojas Satián

Quito, junio de 2026