TRABAJO DE FIN DE GRADO

Grado en Ciencia de Datos

Caracterización y predicción de tipos de lesión en corredores mediante variables biomecánicas, antropométricas y deportivas

Autor: Matías Pérez Lasala Ciacciarelli
Tutor: Héctor Espinós Morato
Curso académico 2025-2026

Agradecimientos

En primer lugar, quiero agradecer a mi tutor, Héctor Espinós Morato, su seguimiento a lo largo de todo el trabajo y su apoyo en los momentos más difíciles. Sin su orientación y ayuda este trabajo no habría sido posible.

Quiero agradecer también a las personas más importantes de mi vida: mi familia. A mi padre, mi madre y mi hermano, por estar ahí en los buenos y en los malos momentos durante todos estos años.

Resumen

Las lesiones musculoesqueléticas son una consecuencia frecuente de la práctica de la carrera. A pesar de décadas de investigación, sigue sin estar claro qué factores determinan qué estructura acaba lesionándose en cada corredor. Buena parte de los estudios disponibles trabajan con muestras pequeñas, generalmente por debajo de los 100 participantes, lo que compromete la estabilidad de sus conclusiones.

Este trabajo examina la asociación entre variables biomecánicas, antropométricas y deportivas con los principales tipos de lesión en 1.402 corredores únicos, a partir de un dataset de evaluación biomecánica en cinta rodante. El diseño es observacional, retrospectivo y transversal. Las lesiones se clasificaron en siete grupos por localización anatómica: rodilla, muslo, pierna baja, cadera/pelvis, pie, tobillo y otras localizaciones.

Los grupos de lesión difieren significativamente en edad, velocidad de carrera y años de experiencia. Los corredores más jóvenes y con menos trayectoria acumulan más lesiones de rodilla; los de mayor edad presentan más lesiones en tobillo, pie y cadera. En el plano biomecánico, nueve de las 80 variables analizadas mostraron diferencias significativas tras corrección por comparaciones múltiples. Las más discriminativas fueron la velocidad pico y la excursión de rotación de rodilla derecha, ambas más elevadas en el grupo de tobillo. Los modelos predictivos alcanzaron una capacidad discriminativa modesta (AUC 0,48–0,61), sin que los métodos de aprendizaje automático superaran de forma relevante a la regresión logística.

Palabras clave: lesiones en corredores, biomecánica de carrera, aprendizaje automático, análisis multivariante, caracterización de lesión.

Abstract

Musculoskeletal injuries are a frequent consequence of running. Despite decades of research, it remains unclear what factors determine which specific structure gets injured in each runner. Much of the available literature relies on small samples, typically below 100 participants, which undermines the stability of their conclusions.

This study examines the association between biomechanical, anthropometric and sport-related variables and the main injury types in 1,402 unique runners, using a dataset of biomechanical evaluation on a treadmill. The design is observational, retrospective and cross-sectional. Injuries were classified into seven groups by anatomical location: knee, thigh, lower leg, hip/pelvis, foot, ankle and other locations.

Injury groups differed significantly in age, running speed and years of experience. Younger, less experienced runners accumulated more knee injuries, while older runners showed higher rates of ankle, foot and hip injuries. At the biomechanical level, nine out of 80 variables analysed showed significant differences after correction for multiple comparisons. The most discriminative were peak velocity and excursion of right knee rotation, both markedly higher in the ankle injury group. Predictive models achieved modest discriminative capacity (AUC 0.48–0.61), with no machine learning method meaningfully outperforming logistic regression.

Keywords: running injuries, running biomechanics, machine learning, multivariate analysis, injury characterisation.

Índice de contenidos

1. Introducción

1.1. La carrera a pie y la incidencia de lesiones
1.2. Principales tipos de lesión y su base biomecánica
1.3. Variables biomecánicas y riesgo de lesión
1.4. Limitaciones de la evidencia disponible
1.5. Aprendizaje automático en lesiones deportivas
1.6. Justificación del presente trabajo

2. Objetivos

2.1. Objetivo general
2.2. Objetivos específicos

3. Marco teórico

4. Material y métodos

4.1. Diseño del estudio
4.2. Muestra y base de datos
4.3. Variables
4.4. Descripción de las variables biomecánicas retenidas
4.5. Definición y operativización de los tipos de lesión
4.6. Análisis estadístico

5. Resultados

5.1. Descripción de la muestra
5.2. Distribución por tipo de lesión y diagnósticos específicos
5.3. Caracterización por tipo de lesión: variables antropométricas y deportivas
5.4. Caracterización por tipo de lesión: variables biomecánicas
5.5. Análisis post-hoc de Dunn
5.6. Modelado explicativo
5.7. Modelado predictivo

6. Discusión

7. Limitaciones

8. Objetivos de Desarrollo Sostenible

9. Conclusiones y líneas futuras

10. Referencias bibliográficas

Anexos

Tabla A1. Variables antropométricas, deportivas y de lesión
Tabla A2. Variables biomecánicas, resumen por dominio
Tabla A4-A9. Post-hoc de Dunn por variable (LEFT_STRIDE_LENGTH, LEFT_KNEE_ADD_PEAK_ANGLE, LEFT_KNEE_ABD_PEAK_ANGLE, RIGHT_ANKLE_EVE_PEAK_VEL, RIGHT_ANKLE_DF_PEAK_ANGLE, LEFT_MHW_EXC_FROM_TO)
Tabla A10. Versiones de librerías del entorno de análisis
Tabla A11. Hiperparámetros de los modelos predictivos
Tabla A12. Estructura de notebooks del repositorio
Declaración del uso de Inteligencia Artificial

Índice de tablas

Tabla 1. Características descriptivas de la muestra
Tabla 2. Distribución por tipo de lesión
Tabla 2b. Top 15 diagnósticos específicos
Tabla 3. Medianas de variables antropométricas y deportivas por tipo de lesión
Tabla 4. Tipos de variables biomecánicas retenidas
Tabla 5. Variables biomecánicas con diferencias significativas
Tabla 6. Medianas de variables biomecánicas significativas por tipo de lesión
Tabla A3. Kruskal-Wallis con FDR para las 80 variables biomecánicas
Tabla 7. Post-hoc de Dunn: RIGHT_KNEE_ROT_PEAK_VEL
Tabla 8. Post-hoc de Dunn: RIGHT_KNEE_ROT_EXCURSION
Tabla 9. Post-hoc de Dunn: RIGHT_STRIDE_LENGTH
Tabla 10. Rendimiento de la regresión logística binaria
Tabla 11. Comparación completa de modelos predictivos
Tabla 12. Análisis de sensibilidad: rendimiento de la regresión logística con y sin la variable años de experiencia corriendo

Índice de figuras

Figura 1. Distribución de la muestra por estado de lesión y tipo anatómico
Figura 2. Comparación de variables demográficas entre lesionados y no lesionados
Figura 3. Top 15 diagnósticos específicos entre los sujetos lesionados
Figura 4. Perfiles biomecánicos por grupo de lesión (heatmap)
Figura 5. Distribución de las principales variables biomecánicas por grupo
Figura 6. Comparación del rendimiento predictivo
Figura 7. Importancia de variables según el random forest
Figura 8. Árbol de decisión para Rodilla vs resto

1. Introducción

1.1. La carrera a pie y la incidencia de lesiones

La carrera a pie se ha convertido en uno de los deportes más practicados a nivel mundial. Su popularidad responde a factores reconocidos: no requiere equipamiento especializado, puede practicarse en entornos muy diversos y sus beneficios para la salud están ampliamente documentados. Sin embargo, esta misma popularidad conlleva una incidencia elevada de lesiones. Las revisiones sistemáticas disponibles estiman que entre el 19% y el 79% de los corredores sufre alguna lesión a lo largo de una temporada, una variabilidad que refleja la heterogeneidad metodológica de los estudios disponibles y la dificultad de establecer definiciones uniformes de lesión (Lopes et al., 2012; van Gent et al., 2007).

La mayor parte de estas lesiones tienen un origen gradual. Se desarrollan de forma progresiva, como consecuencia de repetir miles de veces el mismo ciclo de zancada con algún patrón de movimiento que sobrecarga una estructura concreta. Esto las hace especialmente complicadas de diagnosticar en fases tempranas y dificulta también identificar retrospectivamente qué las desencadenó. La rodilla es la zona más afectada en la mayoría de las series publicadas, acumulando entre el 25% y el 45% de todos los casos (van Gent et al., 2007).

Esta pregunta orienta el presente trabajo: ¿qué variables biomecánicas, antropométricas y deportivas se asocian con cada tipo de lesión, y hasta qué punto son capaces de predecirla? Para abordarla se recurrió al dataset de Clermont et al. (2024), publicado en Nature Scientific Data, con registros de evaluación biomecánica en cinta rodante de más de 1.800 corredores.

El análisis se estructura en dos bloques principales. El primero es descriptivo y comparativo: se caracteriza la muestra y se identifican las variables que difieren significativamente entre los siete grupos de lesión definidos. El segundo es predictivo: se evalúa si esas diferencias son suficientes para construir modelos de clasificación con utilidad clínica real, usando regresión logística, árboles de decisión, random forest y XGBoost.

1.2. Principales tipos de lesión y su base biomecánica

Cada tipo de lesión en corredores presenta un mecanismo etiopatogénico descrito en la literatura, cuyo conocimiento orienta la selección de variables biomecánicas relevantes. El síndrome patelofemoral es el diagnóstico más frecuente. Se manifiesta como dolor en la parte anterior de la rodilla que empeora al bajar escaleras o correr cuesta abajo, y se asocia con una mayor aducción de cadera y valgo dinámico de rodilla durante el apoyo, que aumentan la presión de la rótula sobre el cóndilo femoral (Powers, 2010; Neal et al., 2016). Justo detrás en frecuencia aparece el síndrome de la banda iliotibial, que afecta al compartimento lateral de la rodilla por la fricción repetida sobre el epicóndilo lateral femoral. Estudios prospectivos han identificado la aducción de cadera y la rotación interna de rodilla como los factores biomecánicos más diferenciadores en corredores que acaban desarrollando esta lesión (Noehren et al., 2007).

En la pierna baja, el síndrome de estrés tibial medial aparece por la acumulación de microdaños en el periostio tibial debido a fuerzas de torsión repetitivas, y se asocia con el valgo dinámico de rodilla y la pronación excesiva del pie. En el tobillo, la tendinopatía aquílea es probablemente la lesión con mayor impacto funcional: los períodos de inactividad son largos y las tasas de recurrencia pueden superar el 30% en los primeros dos años. En el pie, la fascitis plantar es la entidad más frecuente, relacionada con la restricción de la dorsiflexión de tobillo y el aumento de tensión sobre la fascia plantar en el despegue.

1.3. Variables biomecánicas y riesgo de lesión

La biomecánica de la carrera analiza cuantitativamente lo que ocurre durante el ciclo de zancada: ángulos articulares, velocidades angulares, excursiones articulares, cadencia, longitud de zancada, tiempos de apoyo y vuelo. La hipótesis detrás de buena parte de la investigación en este campo es que ciertos patrones de movimiento incrementan las cargas mecánicas sobre estructuras específicas y, con el tiempo, acaban produciendo lesiones por uso excesivo.

Dos de los parámetros más estudiados son la cadencia y la longitud de zancada. Heiderscheit et al. (2011) demostraron que aumentar la cadencia un 5-10% reduce la absorción de energía en rodilla hasta en un 34%, lo que abrió la puerta a usar la manipulación de la cadencia como estrategia preventiva. En el plano frontal, la aducción de cadera y el valgo dinámico de rodilla han recibido especial atención: Neal et al. (2016), en una revisión sistemática de 14 estudios prospectivos, identificaron la aducción de cadera como el factor biomecánico con mayor evidencia de asociación con el síndrome patelofemoral.

1.4. Limitaciones de la evidencia disponible

A pesar del volumen creciente de investigación, la evidencia biomecánica sobre lesiones en corredores tiene limitaciones metodológicas importantes que conviene tener presentes. El problema más crítico es el tamaño muestral: la mayoría de estudios trabaja con menos de 100 participantes, lo que condiciona la potencia estadística para detectar efectos pequeños y genera estimaciones poco estables. Una revisión sistemática que incluyó 30 estudios prospectivos con 3.404 corredores concluyó que la literatura disponible no apoya en general el uso de variables biomecánicas como factores de riesgo independientes en corredores no élite, con tamaños del efecto triviales o pequeños (Peterson et al., 2022).

Los sistemas de captura de movimiento, las velocidades de evaluación y los criterios diagnósticos varían considerablemente entre estudios, lo que complica la comparación directa de resultados. A eso se suma el problema del diseño: la mayoría de estudios son transversales, y con ese tipo de diseño no es posible saber si las diferencias cinemáticas encontradas existían antes de la lesión o

aparecieron después, un problema que también afecta a este trabajo y que se discute en el apartado de limitaciones.

1.5. Aprendizaje automático en lesiones deportivas

En la última década ha crecido mucho el interés por aplicar modelos de aprendizaje automático a la predicción de lesiones deportivas. Una revisión sistemática de Van Eetvelde et al. (2021) que analizó 17 estudios encontró rendimientos muy variables, con AUC entre 0,52 y 0,87 dependiendo del deporte, el tipo de lesión y las variables incluidas. Los autores señalaron que los principales puntos débiles de la literatura son la interpretabilidad de los modelos y la calidad de las validaciones, con un uso frecuente de validación sobre el mismo conjunto de entrenamiento, lo que infla artificialmente las métricas. En el contexto concreto de la carrera, los modelos sobre datos biomecánicos tabulares han mostrado una capacidad discriminativa modesta, coherente con la naturaleza multifactorial del riesgo de lesión.

1.6. Justificación del presente trabajo

Este trabajo parte de una base de datos de 1.402 corredores únicos evaluados con un sistema de análisis de movimiento tridimensional, procedente del dataset publicado por Clermont et al. (2024) en Nature Scientific Data. Este tamaño muestral es notablemente superior al de la mayoría de estudios previos en el área, lo que permite aplicar técnicas estadísticas y modelos de aprendizaje automático con suficiente potencia para detectar efectos pequeños.

Desde la perspectiva de la ciencia de datos, la pregunta que guía el trabajo es doble: por un lado, qué variables biomecánicas, antropométricas y deportivas se asocian con los distintos tipos de lesión; por otro, hasta qué punto esas asociaciones son suficientes para construir modelos predictivos con utilidad clínica real. Los resultados, desarrollados en las secciones siguientes, responden a ambas preguntas, aunque no siempre en la dirección que cabría esperar.

2. Objetivos

2.1. Objetivo general

El objetivo de este trabajo es analizar la relación entre las variables biomecánicas de la carrera, las variables antropométricas y las variables deportivas con los principales tipos de lesión en corredores, usando tanto técnicas estadísticas clásicas como modelos de aprendizaje automático.

2.2. Objetivos específicos

Para llegar a ese objetivo general, el trabajo se estructura en torno a varios objetivos más concretos:

- Describir la composición de la muestra y la distribución de los tipos de lesión según localización anatómica, diagnóstico específico, severidad y lateralidad.
- Comparar los siete grupos de lesión en variables antropométricas, deportivas y biomecánicas, e identificar cuáles muestran diferencias significativas tras corregir por comparaciones múltiples.
- Cuantificar el tamaño del efecto de las variables significativas, distinguiendo entre significación estadística y relevancia práctica, dado que ambas dimensiones no siempre coinciden.
- Identificar mediante análisis post-hoc qué pares concretos de grupos de lesión difieren entre sí en cada variable significativa.
- Ajustar modelos de regresión logística binaria 1-vs-resto para los tres grupos principales, obteniendo odds ratios con intervalos de confianza al 95%.
- Evaluar el rendimiento predictivo de árbol de decisión, random forest y XGBoost frente a la regresión logística como punto de partida, mediante validación cruzada estratificada de 5 pliegues.
- Documentar todo el proceso de forma reproducible, justificando cada decisión metodológica para que el análisis pueda verificarse o replicarse.

3. Marco teórico

3.1. Biomecánica de la carrera y ciclo de zancada

El ciclo de zancada va desde el contacto de un pie con el suelo hasta el siguiente contacto del mismo pie. Se divide en dos fases: la de apoyo, durante la cual el pie está en contacto con el suelo y el cuerpo gestiona las fuerzas de impacto, y la de vuelo, en la que ningún pie toca el suelo. A mayor velocidad de carrera, la fase de vuelo se alarga y la de apoyo se acorta. Los parámetros cinemáticos más estudiados en relación con las lesiones incluyen los ángulos articulares (la posición angular de una articulación en un momento dado), las velocidades angulares (la rapidez con que cambia ese ángulo), las excursiones articulares (el rango total de movimiento durante una fase del ciclo) y los parámetros temporoespaciales como la cadencia, la longitud de zancada y la anchura de paso. El análisis tridimensional del movimiento mediante sistemas de captura con marcadores reflectivos cutáneos permite cuantificar la cinemática en los tres planos del espacio: sagital (flexión-extensión), frontal (aducción-abducción) y transversal (rotación interna- externa). Cada plano aporta información complementaria sobre los mecanismos de carga sobre las distintas estructuras del miembro inferior.

3.2. Factores de riesgo biomecánico en lesiones por uso excesivo

Las lesiones en corredores son predominantemente lesiones por uso excesivo, resultado de la acumulación de microdaños en tejidos musculoesqueléticos que no disponen de tiempo suficiente para recuperarse entre sesiones. A diferencia de las lesiones traumáticas agudas, no tienen un único momento de aparición sino que se desarrollan gradualmente a lo largo de semanas o meses. La hipótesis subyacente es que determinados patrones cinemáticos concentran carga mecánica sobre estructuras específicas y que su repetición sistemática genera acumulación de microdaños tisulares. Pero la relación dista de ser directa, la mayoría de revisiones sistemáticas encuentran tamaños del efecto pequeños, lo

que sugiere que la biomecánica es uno de varios factores contribuyentes y no un predictor independiente con capacidad discriminativa suficiente (Peterson et al., 2022). Entre los factores biomecánicos con mayor respaldo empírico destacan la aducción de cadera en el plano frontal, asociada con el síndrome patelofemoral y el síndrome de la banda iliotibial (Neal et al., 2016; Noehren et al., 2007); la cadencia y la longitud de zancada, que modulan las fuerzas de reacción del suelo y la absorción de energía en rodilla (Heiderscheit et al., 2011); y la pronación del pie, relacionada con las lesiones de tobillo y pierna baja por su efecto sobre las fuerzas de torsión tibial.

3.3. Aprendizaje automático aplicado a la predicción de lesiones

Los modelos de aprendizaje automático aprenden patrones a partir de datos de entrenamiento y los aplican para hacer predicciones sobre nuevos casos. En el contexto de las lesiones deportivas, se han aplicado principalmente para clasificar el riesgo de lesión o para identificar qué variables son más informativas para distinguir entre grupos de lesión. Los modelos más utilizados en este campo incluyen la regresión logística, que modela la probabilidad de pertenencia a una clase como función lineal de los predictores y produce coeficientes interpretables en forma de odds ratios; los árboles de decisión, que particionan el espacio de predictores de forma jerárquica y son fácilmente interpretables pero tienden a sobreajustarse; el random forest, un conjunto de árboles entrenados sobre submuestras aleatorias cuyas predicciones se agregan para reducir la varianza (Breiman, 2001); y XGBoost, un método de gradient boosting que construye árboles de forma secuencial corrigiendo los errores del modelo anterior (Chen & Guestrin, 2016).

La evaluación de estos modelos requiere métricas adecuadas al desbalance de clases habitual en datos de lesiones. La balanced accuracy promedia la sensibilidad por clase evitando que las clases mayoritarias dominen la métrica. El área bajo la curva ROC (AUC) mide la capacidad del modelo para separar positivos de negativos independientemente del umbral de clasificación. Valores de AUC cercanos a 0,5 indican capacidad discriminativa próxima al azar; valores superiores a 0,7 se consideran clínicamente útiles en este contexto.

4. Material y métodos

4.1. Diseño del estudio

Se trata de un estudio observacional retrospectivo de corte transversal. El carácter observacional implica que no se realizó ninguna intervención sobre los participantes, sino que los datos fueron recogidos previamente en un contexto clínico. La dimensión retrospectiva indica que el análisis se realizó sobre registros existentes, sin seguimiento prospectivo. Y el corte transversal supone que cada sujeto está representado en un único punto temporal, sin evolución longitudinal

Este tipo de diseño es adecuado para identificar asociaciones entre variables, pero tiene una limitación importante que conviene tener clara desde el principio: no permite establecer relaciones causales ni determinar si los patrones biomecánicos identificados preceden a la lesión o son consecuencia de ella. Esto se discute con más detalle en el apartado de limitaciones.

4.2. Muestra y base de datos

La base de datos utilizada fue publicada por Clermont et al. (2024) en Nature Scientific Data y está disponible públicamente. Se construyó a partir de evaluaciones biomecánicas de carrera en cinta rodante a velocidades estandarizadas, combinadas con información clínica sobre el historial de lesiones recogida mediante autoinforme estructurado. Los datos biomecánicos se capturaron con un sistema de análisis de movimiento tridimensional basado en marcadores reflectivos cutáneos colocados en puntos anatómicos estandarizados del miembro inferior y la pelvis. Los participantes procedían principalmente de clínicas de fisioterapia, centros de evaluación biomecánica y clubs de atletismo, lo que explica la alta proporción de sujetos con historial de lesión en la muestra.

El dataset original contenía 1.832 registros. Para garantizar la independencia estadística de las observaciones, se conservó un único registro por sujeto, el correspondiente a la primera sesión de evaluación disponible cronológicamente. Incluir múltiples registros del mismo sujeto habría introducido pseudorreplicación: el tamaño muestral efectivo quedaría artificialmente inflado y los p-valores serían

menores de lo que deberían. Tras aplicar este criterio, la muestra analítica quedó compuesta por 1.402 sujetos únicos.

4.3. Variables

Variables biomecánicas

Se identificaron 152 variables biomecánicas calculadas a partir de los datos de captura de movimiento, cubriendo la cinemática articular de tobillo, rodilla, cadera y pelvis, así como parámetros temporoespaciales como la longitud y frecuencia de zancada, el tiempo de apoyo y el tiempo de vuelo.

Durante la auditoría de calidad se detectó que 72 de estas variables presentaban valor 0 en la totalidad de las filas, lo que implica varianza nula y ausencia total de capacidad informativa. Estas variables no pudieron calcularse en el proceso de exportación desde el software de análisis biomecánico original. Las 80 variables restantes conformaron el bloque biomecánico analítico, con 40 variables por lado. La Tabla A1 recoge la decisión de inclusión o exclusión de todas las variables antropométricas, deportivas y de lesión, tomada antes de iniciar cualquier análisis. La Tabla A2 resume las variables biomecánicas por dominio anatómico.

Variables antropométricas y deportivas

Se incluyeron como covariables: edad en años, talla en centímetros, peso en kilogramos, velocidad de carrera en metros por segundo, años de experiencia corriendo, género y nivel de práctica deportiva (recreativo o competitivo). Para las variables con posibles valores implausibles se definieron rangos de plausibilidad fisiológica: talla 140–220 cm, peso 40–160 kg, edad 15–80 años, velocidad 1,5–7,0 m/s, y años corriendo no superiores a la edad del sujeto. Los valores fuera de rango se reemplazaron por datos ausentes en versiones limpias de cada variable, sin eliminar filas completas.

La Tabla A1 recoge la decisión de inclusión o exclusión de todas las variables antropométricas, deportivas y de lesión, tomada antes de iniciar cualquier análisis. La Tabla A2 resume las variables biomecánicas por dominio anatómico, incluyendo cuántas se excluyeron por varianza nula en cada grupo.

Variable dependiente

La variable dependiente fue el tipo de lesión según localización anatómica primaria, construida a partir del campo InjJoint del dataset. Las categorías se agruparon en siete grupos con N ≥ 20: rodilla, muslo, pierna baja, cadera y pelvis, pie, tobillo y otras localizaciones. Las categorías con N < 20 se colapsaron en el grupo Other. Los sujetos sin lesión se mantuvieron en el dataset para los análisis binarios pero se excluyeron de los análisis comparativos entre tipos de lesión.

4.4. Descripción de las variables biomecánicas retenidas

La Tabla 4 describe los 40 tipos de variables biomecánicas disponibles en el dataset, cada una en versión izquierda y derecha, sumando las 80 variables del bloque analítico. Su lectura facilita la interpretación de los patrones presentados en los resultados, al precisar qué mide cada variable y en qué dominio articular se encuadra.

Tabla 4. Tipos de variables biomecánicas retenidas (40 tipos × 2 lados = 80 variables)

Tipo de variable	Lado	Unidad	Dominio	Descripción
STRIDE_LENGTH	L/R	m	Temporal-espacial	Longitud de zancada entre apoyos consecutivos del mismo pie
STRIDE_RATE	L/R	pasos/min	Temporal-espacial	Frecuencia de zancada (cadencia de carrera)
SWING_TIME	L/R	s	Temporal-espacial	Tiempo de vuelo por ciclo de zancada
STANCE_TIME	L/R	s	Temporal-espacial	Tiempo de apoyo en tierra por ciclo
STEP_WIDTH	L/R	m	Temporal-espacial	Anchura de paso (distancia mediolateral)
VERTICAL_OSCILLATION	L/R	mm	Temporal-espacial	Oscilación vertical del centro de masas
ANKLE_DF_PEAK_ANGLE	L/R	°	Tobillo sagital	Ángulo pico de dorsiflexión de tobillo en apoyo
ANKLE_EVE_PEAK_ANGLE	L/R	°	Tobillo frontal	Ángulo pico de eversión de tobillo (pronación)
ANKLE_EVE_EXCURSION	L/R	°	Tobillo frontal	Rango total de eversión en el ciclo de apoyo

Tipo de variable	Lado	Unidad	Dominio	Descripción
ANKLE_EVE_PEAK_VEL	L/R	°/s	Tobillo frontal	Velocidad angular máxima de eversión (pronación rápida)
ANKLE_EVE_PERCENT_STANCE	L/R	%	Tobillo frontal	Porcentaje del apoyo en que se alcanza el pico de eversión
ANKLE_ROT_PEAK_ANGLE	L/R	°	Tobillo transversal	Ángulo pico de rotación axial de tobillo
ANKLE_ROT_PEAK_VEL	L/R	°/s	Tobillo transversal	Velocidad angular máxima de rotación de tobillo
ANKLE_ROT_EXCURSION	L/R	°	Tobillo transversal	Rango de rotación de tobillo por ciclo
FOOT_ANG_AT_HS	L/R	°	Pie	Ángulo del pie en el contacto inicial (heel strike)
FOOT_PROG_ANGLE	L/R	°	Pie	Ángulo de progresión del pie durante el apoyo
KNEE_FLEX_PEAK_ANGLE	L/R	°	Rodilla sagital	Ángulo pico de flexión de rodilla en el apoyo
KNEE_ADD_PEAK_ANGLE	L/R	°	Rodilla frontal	Ángulo pico de aducción de rodilla (valgo dinámico)
KNEE_ADD_EXCURSION	L/R	°	Rodilla frontal	Rango de aducción de rodilla en el ciclo de apoyo
KNEE_ADD_PEAK_VEL	L/R	°/s	Rodilla frontal	Velocidad angular máxima de aducción de rodilla
KNEE_ABD_PEAK_ANGLE	L/R	°	Rodilla frontal	Ángulo pico de abducción de rodilla
KNEE_ABD_EXCURSION	L/R	°	Rodilla frontal	Rango de abducción de rodilla por ciclo
KNEE_ABD_PEAK_VEL	L/R	°/s	Rodilla frontal	Velocidad angular máxima de abducción de rodilla
KNEE_ROT_PEAK_ANGLE	L/R	°	Rodilla transversal	Ángulo pico de rotación axial de rodilla
KNEE_ROT_PEAK_VEL	L/R	°/s	Rodilla transversal	Velocidad angular máxima de rotación de rodilla
KNEE_ROT_EXCURSION	L/R	°	Rodilla transversal	Rango total de rotación de rodilla por ciclo de zancada
HIP_EXT_PEAK_ANGLE	L/R	°	Cadera sagital	Ángulo pico de extensión de cadera en el apoyo

Tipo de variable	Lado	Unidad	Dominio	Descripción
HIP_ADD_PEAK_ANGLE	L/R	°	Cadera frontal	Ángulo pico de aducción de cadera
HIP_ADD_EXCURSION	L/R	°	Cadera frontal	Rango de aducción de cadera por ciclo
HIP_ABD_PEAK_VEL	L/R	°/s	Cadera frontal	Velocidad angular máxima de abducción de cadera
HIP_ROT_PEAK_ANGLE	L/R	°	Cadera transversal	Ángulo pico de rotación axial de cadera
HIP_ROT_PEAK_VEL	L/R	°/s	Cadera transversal	Velocidad angular máxima de rotación de cadera
HIP_ROT_EXCURSION	L/R	°	Cadera transversal	Rango total de rotación de cadera por ciclo
PEAK_HIP_ADD_VELOCITY	L/R	°/s	Cadera frontal	Velocidad pico de aducción de cadera (relacionado con valgo de rodilla)
PELVIS_PEAK_DROP_ANGLE	L/R	°	Pelvis	Ángulo pico de caída pélvica contralateral (Trendelenburg)
PELVIS_DROP_EXCURSION	L/R	°	Pelvis	Rango de caída pélvica por ciclo de zancada
PEAK_PELVIC_DROP_VELOCITY	L/R	°/s	Pelvis	Velocidad máxima de caída pélvica
PRONATION_ONSET	L/R	%	Pronación	Inicio de la pronación (% del ciclo de apoyo)
PRONATION_OFFSET	L/R	%	Pronación	Final de la pronación (% del ciclo de apoyo)
MHW_EXC_FROM_TO	L/R	°	Heel whip	Excursión del heel whip medial durante la fase de vuelo

Nota. L/R: cada tipo existe en versión izquierda (LEFT_) y derecha (RIGHT_), sumando 80 variables en total. Las 72 variables excluidas por varianza nula corresponden a versiones de estos mismos tipos que no pudieron calcularse en la exportación desde el software original.

4.5. Definición y operativización de los tipos de lesión

La clasificación de cada sujeto como lesionado o no lesionado se realizó a partir de la combinación coherente de tres campos del dataset: la definición de severidad (InjDefn), la localización articular (InjJoint) y el lado afectado (InjSide). Se consideró no lesionado estricto a quien presentaba simultáneamente una definición de no lesión, ausencia de localización y ausencia de diagnóstico. Los casos con inconsistencias entre campos se identificaron con indicadores de calidad y se trataron de forma conservadora.

La severidad se codificó como variable ordinal con tres niveles: nivel 1, el sujeto continúa entrenando con dolor; nivel 2, el volumen o la intensidad del entrenamiento se ven afectados; nivel 3, el sujeto pierde al menos dos entrenamientos consecutivos. Esta codificación es la empleada en el dataset original (Clermont et al., 2024) y permite capturar un espectro gradual del impacto funcional de la lesión.

4.6. Análisis estadístico

Preparación y auditoría de datos

Antes de realizar cualquier análisis se elaboró una tabla maestra con todas las variables y una decisión de inclusión o exclusión justificada para cada una. Las variables con más del 50% de valores ausentes se excluyeron de los modelos predictivos. Las variables biomecánicas no se imputaron en ningún caso. Para el resto se aplicó eliminación por lista en el modelado, documentando el número de sujetos excluidos en cada paso.

Análisis descriptivo

Las variables continuas se describen mediante mediana e intervalo intercuartílico (IQR), dado que el test de Shapiro-Wilk rechazó la normalidad en todas ellas ($p < 0,001$). Se reportan también la media y la desviación estándar. Las variables categóricas se describen con frecuencias absolutas y relativas.

Análisis comparativo por tipo de lesión

Para comparar los siete grupos de lesión en variables continuas se utilizó el test de Kruskal-Wallis. El tamaño del efecto se estimó mediante eta cuadrado (eta²), con los umbrales habituales: menos de 0,01 efecto negligible, entre 0,01 y 0,06 pequeño, entre 0,06 y 0,14 moderado, y por encima de 0,14 grande. Para variables categóricas se usó chi-cuadrado con V de Cramér. Dado que se realizaron 80 pruebas simultáneas sobre el bloque biomecánico, se aplicó corrección FDR de Benjamini-Hochberg (Benjamini & Hochberg, 1995), considerando significativas las variables con p ajustado < 0,05. Para las variables que superaron ese umbral se realizó el test post-hoc de Dunn (Dunn, 1964) con corrección FDR sobre las 21 comparaciones por pares.

Modelado estadístico y predictivo

Selección de modelos. Se evaluaron cuatro modelos con distinto nivel de complejidad: regresión logística como punto de partida lineal e interpretable, árbol de decisión (max_depth=4) como modelo no lineal simple y fácilmente visualizable, random forest (200 árboles, Breiman, 2001) como método de conjunto que reduce la varianza del árbol individual, y XGBoost (Chen & Guestrin, 2016) como método de gradient boosting que construye árboles de forma secuencial corrigiendo los errores del modelo anterior. Esta progresión permite evaluar si la complejidad creciente del modelo aporta capacidad discriminativa real o si las relaciones entre variables y tipo de lesión son suficientemente lineales para que la regresión logística sea competitiva. No se realizó búsqueda de hiperparámetros: los parámetros se fijaron a valores estándar de la literatura para evitar el sobreajuste al conjunto de validación, dado el tamaño muestral disponible y el número de clases.

Selección de predictores

Los 11 predictores incluidos en el modelado fueron las variables que mostraron diferencias significativas en la fase de caracterización, excluyendo LEFT_STRIDE_LENGTH por correlación bilateral perfecta con RIGHT_STRIDE_LENGTH (r = 1,000). Este procedimiento puede introducir un optimismo moderado: al seleccionar los predictores con información de todos los sujetos, se incluyen también los que actuarán como conjunto de test en cada

pliegue. Dado que los AUC obtenidos son bajos en todos los casos (0,48–0,61), este sesgo no altera las conclusiones principales, pero queda documentado explícitamente como limitación del análisis.

Procedimiento de validación

Se intentó primero un modelo de regresión logística multinomial con los siete grupos simultáneos. Al obtener un rendimiento próximo al nivel de azar, se adoptó la estrategia de modelos binarios 1-vs-resto para los tres grupos principales. Todos los predictores se estandarizaron mediante z-score dentro de cada pliegue de validación cruzada usando un pipeline de sklearn, de forma que la estandarización se ajusta únicamente sobre el conjunto de entrenamiento y se aplica al de test, evitando el data leakage. La evaluación se realizó mediante validación cruzada estratificada de 5 pliegues, reportando balanced accuracy, F1, AUC, sensibilidad y especificidad. Todos los modelos incluyeron ajuste para el desbalance de clases. Todos los análisis se realizaron en Python 3.9 con semilla fija (random_state = 42).

Análisis de sensibilidad por valores ausentes

La variable años de experiencia corriendo presentó un 25,0% de valores ausentes, lo que redujo la muestra de modelado de 978 a 755 sujetos lesionados al aplicar eliminación por lista. Para evaluar si esta reducción afecta a los resultados, se comparó el rendimiento de los modelos en dos configuraciones: con los 11 predictores originales (N = 755) y excluyendo la variable años corriendo para disponer de la muestra completa de lesionados (N = 979). Los resultados de esta comparación se presentan en el apartado de modelado predictivo (Tabla 12). La variable se mantuvo en el modelo principal porque su aportación informativa, aunque modesta, está respaldada por la literatura sobre factores de riesgo en corredores, y porque su exclusión produjo mejoras marginales y consistentes en los tres grupos (+0,011 a +0,024 en AUC), pero de magnitud insuficiente para alterar las conclusiones principales.

Las versiones exactas de las librerías empleadas y los hiperparámetros de cada modelo se recogen en las Tablas A10 y A11 del anexo, de forma que el análisis pueda reproducirse exactamente.

5. Resultados

5.1. Descripción de la muestra

La muestra analítica estuvo compuesta por 1.402 sujetos. El 53,6% eran mujeres (n = 752) y el 46,3% hombres (n = 649), con un registro sin dato de género. La mediana de edad fue de 39,0 años (IQR 30,0–48,0), lo que sitúa a la muestra en una franja de corredores adultos de mediana edad, un perfil representativo del corredor recreativo adulto. La talla mediana fue de 172,0 cm (IQR 165,1–179,0) y el peso mediano de 69,1 kg (IQR 60,4–79,3). La velocidad de carrera presentó una mediana de 2,7 m/s (IQR 2,4–2,8), equivalente a un ritmo aproximado de 6 min/km. En cuanto a la experiencia deportiva, la mediana de años corriendo fue de 8,0 años (IQR 3,0–15,0), aunque esta variable presentó un 25,0% de valores ausentes, lo que hay que tener en cuenta al interpretar los resultados que la involucran. El 65,6% practicaba la carrera a nivel recreativo y el 25,5% a nivel competitivo.

Se detectaron valores implausibles en un número reducido de registros: seis casos con talla fuera del rango de plausibilidad (0,4%), cinco con peso implausible (0,4%), un caso con edad superior a 80 años (0,1%) y 51 registros con años corriendo superiores a la edad del sujeto (3,6%). En todos los casos se crearon versiones limpias de las variables afectadas, conservando el resto de información del registro.

Tabla 1. Características descriptivas de la muestra (N = 1.402)

Variable	N válido	Mediana (IQR)	Media	DE
Edad (años)	1.401	39,0 (30,0–48,0)	39,4	12,0
Talla (cm)	1.396	172,0 (165,1–179,0)	172,1	9,0
Peso (kg)	1.397	69,1 (60,4–79,3)	70,7	13,0
Velocidad de carrera (m/s)	1.402	2,7 (2,4–2,8)	2,6	0,4
Años corriendo	1.000	8,0 (3,0–15,0)	10,8	9,7

Nota. IQR: intervalo intercuartílico. DE: desviación estándar.

5.2. Distribución por tipo de lesión y diagnósticos específicos

Del total de 1.402 sujetos, el 69,8% presentaba algún tipo de lesión registrada (n = 979) y el 30,2% no (n = 423). Esta proporción es elevada, pero responde a una característica conocida de la fuente de datos: la muestra procede principalmente de clínicas de fisioterapia y centros de evaluación biomecánica, entornos a los que los corredores acuden con frecuencia por motivo de lesión. No debe interpretarse como prevalencia poblacional de lesiones en corredores en general.

La rodilla fue la localización más frecuente, concentrando el 33,2% de los casos (n = 325), seguida del muslo (14,6%, n = 143), la pierna baja (14,4%, n = 141), la cadera y pelvis (11,2%, n = 110), el pie (11,0%, n = 108), el tobillo (8,4%, n = 82) y otras localizaciones (7,2%, n = 70). En cuanto a la severidad, el nivel 2 fue el más frecuente (42,0%), seguido del nivel 1 (27,6%) y el nivel 3 (22,4%). El 8,1% restante correspondía a registros sin dato de severidad. La distribución por lateralidad fue bastante equilibrada: 32,9% lado dominante, 30,8% no dominante y 25,0% bilateral. El 11,2% restante no contaba con dato de lateralidad registrado.

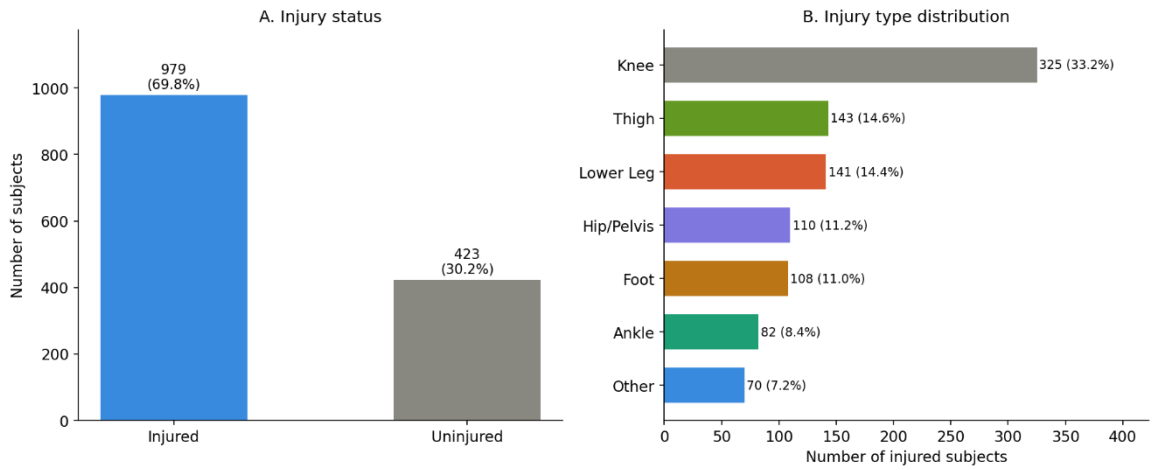


Figura 1. Distribución de la muestra por estado de lesión (A) y por tipo de lesión anatómica entre los sujetos lesionados (B).

Tabla 2. Distribución por tipo de lesión (N lesionados = 979)

Tipo de lesión	N	% lesionados	Diagnósticos más frecuentes
Rodilla	325	33,2%	Sínd. patelofemoral (12,5%), sínd. banda iliotibial (10,0%)
Muslo	143	14,6%	Distensión isquiotibial (2,9%)
Pierna baja	141	14,4%	Distensión gemelar (4,3%), SDMT (3,1%), shin splints (1,9%)
Cadera/pelvis	110	11,2%	Bursitis (1,9%), tendinitis de cadera
Pie	108	11,0%	Fascitis plantar (4,6%)
Tobillo	82	8,4%	Tendinitis de Aquiles (4,9%)
Otras localizaciones	70	7,2%	Diversas (dolor lumbar 0,8%, fractura de estrés 0,7%)

Nota. SDMT: síndrome de estrés tibial medial. Los porcentajes de diagnósticos están calculados sobre el total de lesionados (N = 979).

Tabla 2b. Top 15 diagnósticos específicos entre los sujetos lesionados (N = 979)

Diagnóstico específico	N	% sobre lesionados
Síndrome patelofemoral	122	12,5%
Síndrome de la banda iliotibial	98	10,0%
Tendinitis de Aquiles	48	4,9%
Fascitis plantar	45	4,6%
Distensión gemelar	42	4,3%
Síndrome de estrés tibial medial	30	3,1%
Distensión isquiotibial	28	2,9%
Bursitis	19	1,9%
Shin splints	19	1,9%
Osteoartritis	12	1,2%
Tendinitis (inespecífica)	10	1,0%
Dolor lumbar	8	0,8%
Fractura de estrés tibial	7	0,7%
Distensión de flexor de cadera	7	0,7%
Otros diagnósticos específicos	≥ 5 c/u	—

Nota. Se excluyen 9 registros con diagnóstico 'fill in specifics below' por información incompleta.

Figura 3. Top 15 diagnósticos específicos entre los sujetos lesionados

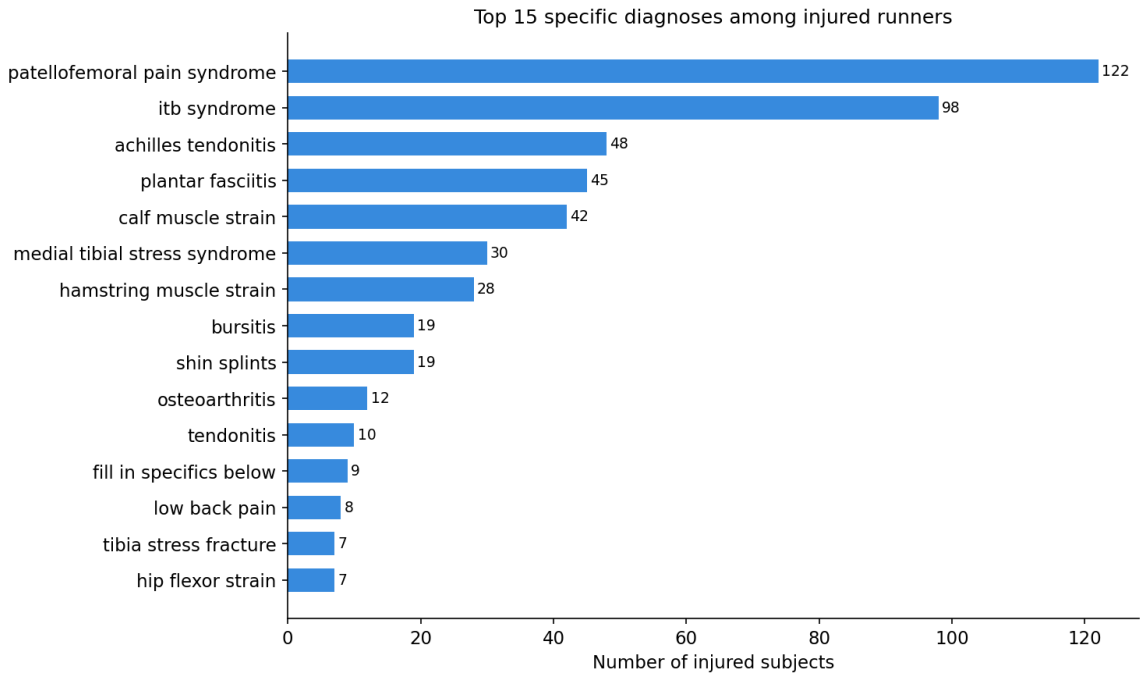


Figura 3. Top 15 diagnósticos específicos entre los sujetos lesionados (N = 979). Se excluyen 9 registros con diagnóstico 'fill in specifics below'.

5.3. Caracterización por tipo de lesión: Variables antropométricas y deportivas

La comparación de los siete grupos mediante el test de Kruskal-Wallis reveló diferencias estadísticamente significativas en tres variables tras la corrección FDR: edad ($H = 38,2$; p ajustado < 0,001; $eta^2 = 0,033$), velocidad de carrera ($H = 27,8$; p ajustado = 0,0003; $eta^2 = 0,022$) y años de experiencia corriendo ($H = 14,7$; p ajustado = 0,037; $eta^2 = 0,012$). Talla y peso no mostraron diferencias significativas, ni tampoco el género ni el nivel competitivo.

Cabe señalar que la velocidad de carrera, aunque estadísticamente significativa, debe interpretarse con cautela en el contexto de este análisis comparativo. Al tratarse de una variable que condiciona directamente muchos parámetros biomecánicos: longitud de zancada, tiempos de apoyo y vuelo, velocidades angulares, su variación entre grupos de lesión puede actuar como factor de confusión en la interpretación de las diferencias cinemáticas observadas en secciones posteriores.

Es importante matizar que los tres tamaños del efecto son pequeños, con eta^2 entre 0,012 y 0,033. Hay diferencias reales entre grupos, pero el solapamiento es considerable y no permiten clasificar a un corredor en un grupo de lesión a partir de estas variables por sí solas.

Tabla 3. Medianas (IQR) de las variables antropométricas y deportivas significativas por tipo de lesión

Tipo lesión	Edad (años)	Velocidad (m/s)	Años corriendo	N edad	N vel.	N años
Ankle	42,0 (31,0–50,5)	2,7 (2,4–2,9)	9,5 (5,0–17,0)	82	82	68
Foot	43,5 (37,0–51,2)	2,5 (2,2–2,8)	9,0 (4,0–16,0)	108	108	89
Hip/Pelvis	41,0 (34,0–49,0)	2,6 (2,3–2,8)	9,5 (4,0–16,0)	109	110	94
Knee	37,0 (29,0–45,0)	2,6 (2,4–2,7)	6,0 (3,0–13,0)	325	325	233
Lower Leg	39,0 (33,0–48,0)	2,7 (2,3–2,9)	7,0 (3,0–14,5)	141	141	111
Other	43,0 (38,0–47,5)	2,4 (2,2–2,7)	10,0 (4,5–21,0)	70	70	51

Tipo lesión	Edad (años)	Velocidad (m/s)	Años corriendo	N edad	N vel.	N años
Thigh	37,0 (29,5–45,0)	2,7 (2,5–2,8)	6,5 (3,0–15,0)	143	143	110

Nota. Las N de 'Años corriendo' son inferiores por los valores ausentes en la variable (25% en la muestra total).

Este patrón es consistente con lo descrito en la literatura clínica. Los grupos de pie y tobillo presentaron las medianas de edad más altas (43,5 y 42,0 años respectivamente), mientras que rodilla y muslo mostraron las más bajas (37,0 años en ambos). Los corredores más jóvenes presentan mayor concentración de lesiones en rodilla, mientras que los de mayor edad y experiencia acumulan más lesiones en estructuras distales como el tobillo o la cadera, lo que puede relacionarse con la mayor carga mecánica acumulada sobre esas estructuras a lo largo de años de práctica. Para los años de experiencia se observa el mismo patrón: rodilla y muslo tienen las medianas más bajas (6,0 y 6,5 años), frente a tobillo y cadera/pelvis (9,5 años ambos).

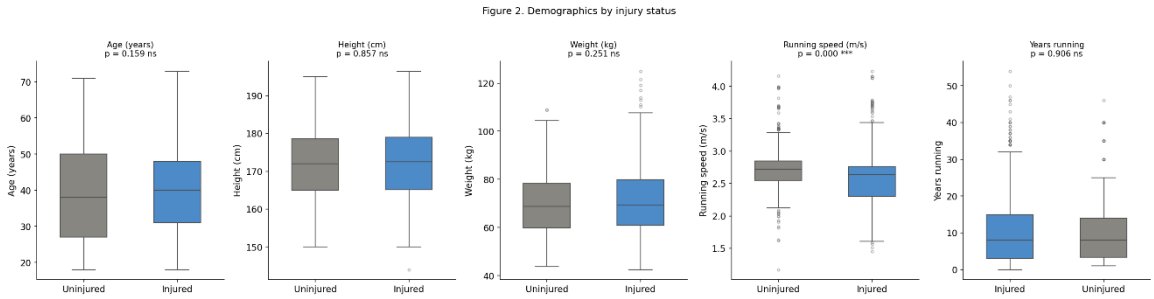


*Figura 2. Comparación de variables demográficas entre sujetos lesionados y no lesionados. *** p < 0,001; ns = no significativo (Mann-Whitney U).*

5.4. Caracterización por tipo de lesión: Variables biomecánicas

De las 80 variables biomecánicas testadas, 9 mostraron diferencias estadísticamente significativas entre los siete grupos de lesión tras la corrección FDR (p ajustado < 0,05). Los tamaños del efecto resultaron modestos en todos los casos, con eta² entre 0,013 y 0,028, coherentes con lo que reportan revisiones previas sobre biomecánica de la carrera y lesiones (Peterson et al., 2022; Van der Worp et al., 2016).

Tabla 5. Variables biomecánicas con diferencias significativas entre grupos de lesión (Kruskal-Wallis con FDR)

Variable	H	p crudo	p FDR	eta²
RIGHT_KNEE_ROT_PEAK_VEL	32,70	< 0,001	0,001	0,028
RIGHT_KNEE_ROT_EXCURSION	30,74	< 0,001	0,001	0,025
RIGHT_STRIDE_LENGTH	28,09	< 0,001	0,002	0,023
LEFT_STRIDE_LENGTH	27,93	< 0,001	0,002	0,023
LEFT_KNEE_ADD_PEAK_ANGLE	24,01	< 0,001	0,008	0,019
LEFT_KNEE_ABD_PEAK_ANGLE	20,14	0,003	0,035	0,015
RIGHT_ANKLE_EVE_PEAK_VEL	19,15	0,004	0,041	0,014
RIGHT_ANKLE_DF_PEAK_ANGLE	19,05	0,004	0,041	0,013
LEFT_MHW_EXC_FROM_TO	18,48	0,005	0,046	0,013

Nota. FDR: False Discovery Rate (Benjamini-Hochberg, 1995). eta²: eta cuadrado. Solo se muestran las 9 variables con p ajustado < 0,05.

Tabla 6. Medianas (IQR) de las variables biomecánicas significativas por tipo de lesión

Variable	Ankle	Foot	Hip/Pelvis	Knee	Lower Leg	Other	Thigh
R_KNEE_ROT_PEAK_VEL (°/s)	224,0 (182,9–260,2)	223,2 (171,8–269,5)	187,4 (152,3–224,7)	184,6 (147,2–231,7)	193,0 (157,1–234,7)	197,3 (155,6–269,0)	198,1 (152,6–239,3)
R_KNEE_ROT_EXCURSION (°)	15,56 (12,91–18,29)	15,37 (12,86–18,49)	13,56 (10,97–16,36)	13,09 (10,52–16,84)	13,73 (11,59–16,95)	14,70 (11,73–18,44)	14,29 (11,08–17,36)
R_STRIDE_LENGTH (m)	1,97 (1,73–2,09)	1,79 (1,61–2,03)	1,86 (1,71–2,02)	1,90 (1,73–2,03)	1,93 (1,73–2,09)	1,78 (1,63–1,90)	1,96 (1,78–2,06)
L_KNEE_ADD_PEAK (°)	–4,68 (–7,3–1,6)	–5,25 (–7,8–2,3)	–4,62 (–8,2–1,7)	–4,77 (–6,8–1,8)	–3,01 (–5,7–1,0)	–4,65 (–8,0–1,5)	–3,21 (–6,0–0,7)

Variable	Ankle	Foot	Hip/Pelvis	Knee	Lower Leg	Other	Thigh
L_KNEE_ABD_PEAK (°)	–10,15 (–13,8–7,2)	–11,09 (–14,2–8,0)	–10,83 (–13,9–7,9)	–9,53 (–12,8–7,3)	–8,87 (–11,9–5,9)	–10,94 (–14,0–6,8)	–9,43 (–11,8–6,6)
R_ANKLE_EVE_PEAK_VEL (°/s)	–230 (–298–172)	–219 (–257–166)	–216 (–260–166)	–202 (–248–155)	–222 (–273–176)	–196 (–239–161)	–223 (–273–176)
R_ANKLE_DF_PEAK (°)	20,26 (17,78–23,06)	20,55 (18,34–22,44)	20,11 (17,78–22,63)	21,55 (19,24–23,56)	21,61 (17,97–24,33)	20,82 (18,57–22,60)	21,44 (19,28–23,41)
L_MHW_EXC_FROM_TO (°)	–10,69 (–15,3–7,5)	–10,97 (–14,4–7,0)	–11,08 (–15,3–6,9)	–9,30 (–13,9–6,2)	–8,39 (–12,8–5,3)	–9,48 (–11,7–5,6)	–10,25 (–14,6–7,0)

Nota. R_: RIGHT_; L_: LEFT_. Velocidades en °/s, ángulos en °, longitudes en m. LEFT_STRIDE_LENGTH se omite por correlación bilateral perfecta con RIGHT_STRIDE_LENGTH (r = 1,000). Valores negativos indican dirección de movimiento según la convención del dataset.

La tabla A3 recoge los resultados completos del test de Kruskal-Wallis para las 80 variables biomecánicas, ordenadas de mayor a menor estadístico H. Se incluye aquí por completitud y reproducibilidad: permite ver no solo qué variables fueron significativas, sino también cuánto se quedaron fuera del umbral las que no lo superaron.

Tabla A3. Kruskal-Wallis con FDR para las 80 variables biomecánicas (N = 979 lesionados, 7 grupos)

Variable	H	p crudo	p FDR	eta²	Sig.
RIGHT_KNEE_ROT_PEAK_VEL	32,703	< 0,001	0,001	0,028	Sí
RIGHT_KNEE_ROT_EXCURSION	30,737	< 0,001	0,001	0,025	Sí
RIGHT_STRIDE_LENGTH	28,094	< 0,001	0,002	0,023	Sí
LEFT_STRIDE_LENGTH	27,932	< 0,001	0,002	0,023	Sí
LEFT_KNEE_ADD_PEAK_ANGLE	24,012	< 0,001	0,008	0,019	Sí
LEFT_KNEE_ABD_PEAK_ANGLE	20,141	0,003	0,035	0,015	Sí
RIGHT_ANKLE_EVE_PEAK_VEL	19,153	0,004	0,041	0,014	Sí
RIGHT_ANKLE_DF_PEAK_ANGLE	19,051	0,004	0,041	0,013	Sí
LEFT_MHW_EXC_FROM_TO	18,483	0,005	0,046	0,013	Sí
RIGHT_ANKLE_EVE_EXCURSION	17,464	0,008	0,062	0,012	No
LEFT_KNEE_ADD_EXCURSION	16,683	0,011	0,077	0,011	No
RIGHT_HIP_ROT_PEAK_ANGLE	16,350	0,012	0,079	0,011	No
LEFT_FOOT_PROG_ANGLE	16,103	0,013	0,079	0,010	No

Variable	H	p crudo	p FDR	eta²	Sig.
LEFT_KNEE_ROT_PEAK_VEL	15,980	0,014	0,079	0,010	No
LEFT_KNEE_ROT_EXCURSION	15,810	0,015	0,079	0,010	No
RIGHT_HIP_ROT_EXCURSION	14,656	0,023	0,116	0,009	No
RIGHT_STRIDE_RATE	14,085	0,029	0,130	0,008	No
RIGHT_STANCE_TIME	13,983	0,030	0,130	0,008	No
LEFT_ANKLE_DF_PEAK_ANGLE	13,778	0,032	0,130	0,008	No
LEFT_STANCE_TIME	13,759	0,032	0,130	0,008	No
LEFT_STRIDE_RATE	13,503	0,036	0,134	0,008	No
LEFT_FOOT_ANG_AT_HS	13,408	0,037	0,134	0,008	No
LEFT_ANKLE_ROT_PEAK_ANGLE	13,308	0,038	0,134	0,008	No
RIGHT_HIP_EXT_PEAK_ANGLE	13,184	0,040	0,134	0,007	No
RIGHT_KNEE_ADD_PEAK_ANGLE	12,631	0,049	0,153	0,007	No
LEFT_ANKLE_EVE_EXCURSION	12,616	0,050	0,153	0,007	No
LEFT_HIP_ROT_PEAK_VEL	12,439	0,053	0,157	0,007	No
LEFT_ANKLE_EVE_PEAK_VEL	12,093	0,060	0,162	0,006	No
RIGHT_FOOT_PROG_ANGLE	12,133	0,059	0,162	0,006	No
LEFT_HIP_EXT_PEAK_ANGLE	12,052	0,061	0,162	0,006	No
RIGHT_ANKLE_ROT_PEAK_ANGLE	11,654	0,070	0,179	0,006	No
RIGHT_KNEE_ABD_PEAK_VEL	11,590	0,072	0,179	0,006	No
LEFT_ANKLE_ROT_PEAK_VEL	10,972	0,089	0,214	0,005	No
RIGHT_KNEE_FLEX_PEAK_ANGLE	10,912	0,091	0,214	0,005	No
RIGHT_FOOT_ANG_AT_HS	10,570	0,103	0,228	0,005	No
RIGHT_KNEE_ABD_PEAK_ANGLE	10,601	0,102	0,228	0,005	No
LEFT_HIP_ROT_EXCURSION	10,284	0,113	0,244	0,004	No
RIGHT_HIP_ADD_PEAK_ANGLE	10,063	0,122	0,244	0,004	No
LEFT_KNEE_ABD_PEAK_VEL	10,086	0,121	0,244	0,004	No
RIGHT_PRONATION_ONSET	10,127	0,119	0,244	0,004	No
RIGHT_PEAK_HIP_ADD_VELOCITY	9,813	0,133	0,259	0,004	No
RIGHT_KNEE_ROT_PEAK_ANGLE	9,720	0,137	0,261	0,004	No
RIGHT_KNEE_ADD_EXCURSION	9,631	0,141	0,263	0,004	No
LEFT_HIP_ROT_PEAK_ANGLE	9,518	0,147	0,266	0,004	No
RIGHT_ANKLE_ROT_PEAK_VEL	9,318	0,157	0,278	0,003	No
RIGHT_ANKLE_EVE_PERCENT_STANCE	8,716	0,190	0,331	0,003	No
LEFT_PELVIS_DROP_EXCURSION	8,401	0,210	0,355	0,003	No
LEFT_SWING_TIME	8,289	0,218	0,355	0,002	No
LEFT_PRONATION_OFFSET	8,231	0,222	0,355	0,002	No

Variable	H	p crudo	p FDR	eta²	Sig.
LEFT_ANKLE_EVE_PERCENT_STANCE	8,254	0,220	0,355	0,002	No
LEFT_KNEE_ABD_EXCURSION	8,062	0,234	0,358	0,002	No
LEFT_PELVIS_PEAK_DROP_ANGLE	8,014	0,237	0,358	0,002	No
LEFT_HIP_ADD_PEAK_ANGLE	8,055	0,234	0,358	0,002	No
RIGHT_MHW_EXC_FROM_TO	7,574	0,271	0,401	0,002	No
LEFT_VERTICAL_OSCILLATION	7,282	0,296	0,430	0,001	No
RIGHT_PELVIS_DROP_EXCURSION	7,040	0,317	0,453	0,001	No
RIGHT_SWING_TIME	6,931	0,327	0,459	0,001	No
LEFT_KNEE_FLEX_PEAK_ANGLE	6,758	0,344	0,466	0,001	No
RIGHT_HIP_ADD_EXCURSION	6,803	0,340	0,466	0,001	No
RIGHT_HIP_ABD_PEAK_VEL	6,700	0,350	0,466	0,001	No
RIGHT_PRONATION_OFFSET	6,477	0,372	0,488	0,001	No
RIGHT_KNEE_ADD_PEAK_VEL	6,412	0,379	0,489	0,000	No
RIGHT_VERTICAL_OSCILLATION	6,192	0,402	0,511	0,000	No
LEFT_HIP_ABD_PEAK_VEL	5,632	0,466	0,564	0,000	No
LEFT_HIP_ADD_EXCURSION	5,688	0,459	0,564	0,000	No
RIGHT_PELVIS_PEAK_DROP_ANGLE	5,533	0,478	0,570	0,000	No
LEFT_ANKLE_EVE_PEAK_ANGLE	5,030	0,540	0,617	0,000	No
LEFT_STEP_WIDTH	5,176	0,521	0,605	0,000	No
RIGHT_STEP_WIDTH	5,176	0,521	0,605	0,000	No
LEFT_PRONATION_ONSET	5,923	0,432	0,540	0,000	No
LEFT_KNEE_ADD_PEAK_VEL	4,256	0,642	0,704	0,000	No
LEFT_PEAK_HIP_ADD_VELOCITY	4,478	0,612	0,680	0,000	No
LEFT_KNEE_ROT_PEAK_ANGLE	4,948	0,551	0,620	0,000	No
LEFT_ANKLE_ROT_EXCURSION	3,314	0,769	0,804	0,000	No
LEFT_PEAK_PELVIC_DROP_VELOCITY	3,799	0,704	0,761	0,000	No
RIGHT_HIP_ROT_PEAK_VEL	3,675	0,721	0,769	0,000	No
RIGHT_KNEE_ABD_EXCURSION	3,277	0,773	0,804	0,000	No
RIGHT_ANKLE_ROT_EXCURSION	2,448	0,874	0,885	0,000	No
RIGHT_ANKLE_EVE_PEAK_ANGLE	2,797	0,834	0,855	0,000	No
RIGHT_PEAK_PELVIC_DROP_VELOCITY	1,342	0,969	0,969	0,000	No

Nota. FDR: False Discovery Rate (Benjamini-Hochberg). ✓ = p ajustado < 0,05. Ordenadas de mayor a menor H.

Las dos variables más discriminativas fueron la velocidad pico de rotación de rodilla derecha (eta² = 0,028) y la excursión de rotación de rodilla derecha (eta² = 0,025), seguidas de la longitud de zancada bilateral (eta² ≈ 0,023 en ambos lados). Los grupos Ankle y Foot presentaron los valores más altos de velocidad y excursión de rotación de rodilla, mientras que Knee e Hip/Pelvis los más bajos. Para la longitud de zancada, el grupo Other presentó las zancadas más cortas, diferenciándose de Ankle, Knee, Lower Leg y Thigh. Se detectó correlación bilateral perfecta entre las longitudes de zancada izquierda y derecha (r = 1,000), por lo que para el modelado se utilizó únicamente la versión derecha.

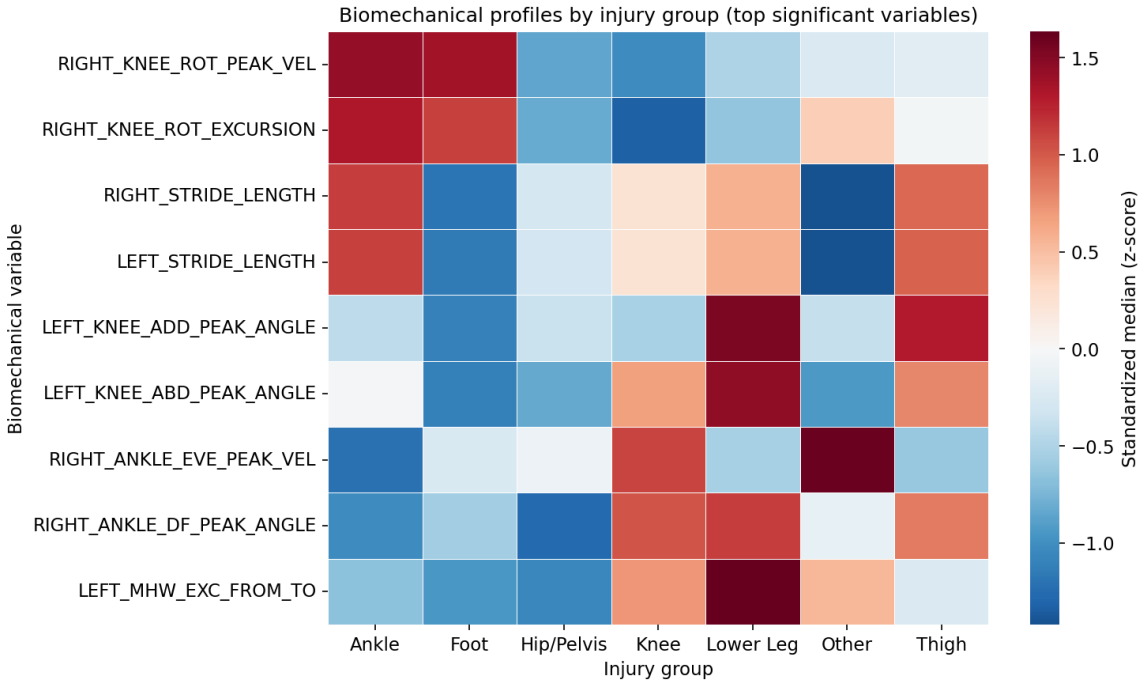


Figura 4. Perfiles biomecánicos por grupo de lesión. Medianas estandarizadas (z-score). Rojo: valores superiores a la media; azul: inferiores.

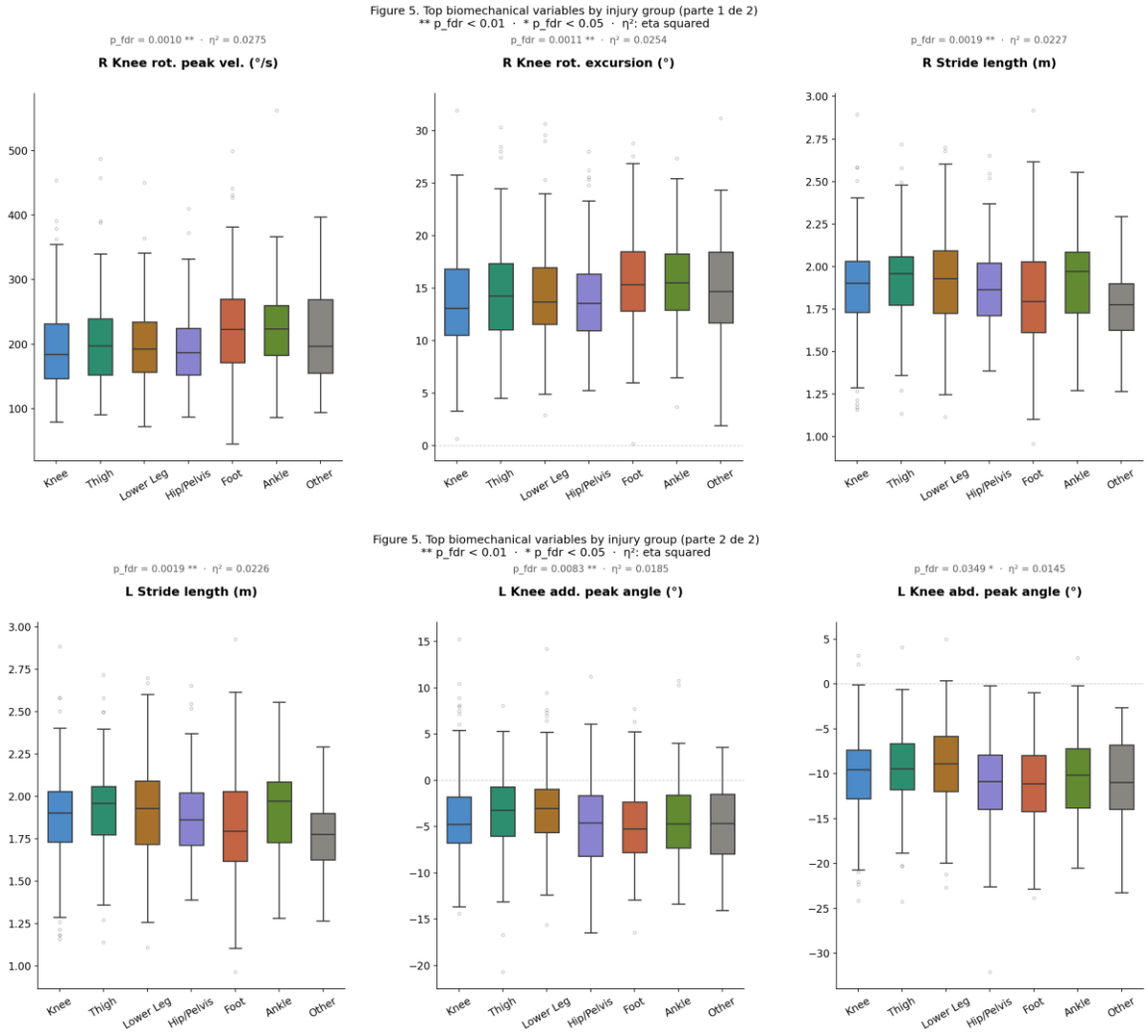


*Figura 5. Distribución de las principales variables biomecánicas por grupo de lesión. ** p ajustado < 0,01; * p ajustado < 0,05. η²: eta cuadrado.*

5.5. Análisis post-hoc de Dunn

Para las 9 variables que superaron la corrección FDR se realizó el test post-hoc de Dunn con corrección FDR sobre las 21 comparaciones por pares posibles entre los siete grupos. El código de colores en las tablas indica el nivel de significación: azul oscuro (*** p < 0,001), azul medio (** p < 0,01), azul claro (* p < 0,05) y sin color (ns, p ≥ 0,05).

Las matrices post-hoc para las seis variables restantes se recogen en las Tablas A4–A9 del anexo, junto con su interpretación individual.

Tabla 7. Post-hoc de Dunn: RIGHT_KNEE_ROT_PEAK_VEL (velocidad pico de rotación de rodilla derecha, °/s)

Variable más discriminativa (eta² = 0,028). El grupo Ankle presenta los valores más altos (mediana 224,0 °/s) y se diferencia significativamente de Knee (p = 0,0003), Hip/Pelvis (p = 0,0014), Lower Leg (p = 0,0051) y Thigh (p = 0,0051). El grupo Foot muestra un patrón similar. Los grupos Knee, Hip/Pelvis, Lower Leg y Thigh no difieren entre sí.

	Ankle	Foot	Hip/Pelvis	Knee	Lower Leg	Other	Thigh
Ankle	—	0,5208	**0,0014****	**< 0,001*****	**0,0051****	0,1685	**0,0051****
Foot	0,5208	—	**0,0051****	**< 0,001*****	**0,0222***	0,4137	**0,0222***
Hip/Pelvis	**0,0014****	**0,0051****	—	0,9786	0,5207	0,1685	0,5207
Knee	**< 0,001*****	**< 0,001*****	0,9786	—	0,4604	0,1179	0,4604
Lower Leg	**0,0051****	**0,0222***	0,5207	0,4604	—	0,4137	0,9786
Other	0,1685	0,4137	0,1685	0,1179	0,4137	—	0,4137
Thigh	**0,0051****	**0,0222***	0,5207	0,4604	0,9786	0,4137	—

*Nota. *** p < 0,001; ** p < 0,01; * p < 0,05; Sin color: ns (p ≥ 0,05).*

Tabla 8. Post-hoc de Dunn: RIGHT_KNEE_ROT_EXCURSION (excursión de rotación de rodilla derecha, °)

Segunda variable más discriminativa (eta² = 0,025). Patrón coherente con la velocidad angular: Ankle (15,56°) y Foot (15,37°) presentan excursiones significativamente mayores que Knee (13,09°) e Hip/Pelvis (13,56°). La coherencia entre velocidad y excursión de rotación refuerza la hipótesis de que los corredores con lesión de tobillo presentan un patrón cinemático de rodilla cualitativamente distinto.

	Ankle	Foot	Hip/Pelvis	Knee	Lower Leg	Other	Thigh
Ankle	—	0,6608	**0,0179***	**< 0,001*****	**0,0284***	0,2814	**0,0490***
Foot	0,6608	—	**0,0284***	**< 0,001*****	0,0506	0,4686	0,0840
Hip/Pelvis	**0,0179***	**0,0284***	—	0,5043	0,6608	0,2663	0,5043
Knee	**< 0,001*****	**< 0,001*****	0,5043	—	0,2663	0,0527	0,1309
Lower Leg	**0,0284***	0,0506	0,6608	0,2663	—	0,4220	0,7475
Other	0,2814	0,4686	0,2663	0,0527	0,4220	—	0,5043
Thigh	**0,0490***	0,0840	0,5043	0,1309	0,7475	0,5043	—

*Nota. *** p < 0,001; ** p < 0,01; * p < 0,05; Sin color: ns.*

Tabla 9. Post-hoc de Dunn: RIGHT_STRIDE_LENGTH (longitud de zancada derecha, m)

El patrón es diferente al de las variables de rotación: el grupo más diferenciado no es Ankle sino Other (otras localizaciones), con las zancadas más cortas (mediana 1,78 m). Other se diferencia significativamente de Ankle (p = 0,001), Knee (p = 0,007), Lower Leg (p = 0,003) y Thigh (p = 0,001). Ankle y Thigh presentan las zancadas más largas.

	Ankle	Foot	Hip/Pelvis	Knee	Lower Leg	Other	Thigh
Ankle	—	**0,0077****	0,0957	0,1577	0,4874	**0,0011****	0,7409
Foot	**0,0077****	—	0,3233	0,0728	**0,0257***	0,3233	**0,0066****
Hip/Pelvis	0,0957	0,3233	—	0,5039	0,2386	0,0678	0,1003
Knee	0,1577	0,0728	0,5039	—	0,4159	**0,0066****	0,1664
Lower Leg	0,4874	**0,0257***	0,2386	0,4159	—	**0,0026****	0,6026
Other	**0,0011****	0,3233	0,0678	**0,0066****	**0,0026****	—	**0,0011****
Thigh	0,7409	**0,0066****	0,1003	0,1664	0,6026	**0,0011****	—

*Nota. *** p < 0,001; ** p < 0,01; * p < 0,05; Sin color: ns.*

En conjunto, el análisis post-hoc confirma que el grupo Ankle es el más diferenciado del resto, especialmente en las variables de rotación de rodilla. El cluster formado por Lower Leg y Thigh, con menor aducción de rodilla izquierda, y el patrón simétrico de la velocidad de eversión de tobillo derecho son los hallazgos más consistentes entre el resto de variables.

5.6. Modelado explicativo

Los 11 predictores seleccionados conformaron el conjunto de variables para el modelado. Tras eliminar por lista los sujetos con valores ausentes, principalmente por la variable años de experiencia corriendo con un 25% de missings, la muestra de modelado quedó compuesta por 755 sujetos lesionados, lo que supone retener el 77,1% del total de 979.

El modelo de regresión logística multinomial con los siete grupos simultáneos obtuvo un balanced accuracy medio de 0,160 (DE = 0,019) en la validación cruzada de 5 pliegues. A modo de referencia, el nivel de azar teórico para siete clases es 0,143, lo que supone una diferencia de apenas 0,017 puntos. El modelo prácticamente no discrimina. Esto llevó a adoptar la estrategia de modelos binarios 1-vs-resto para los tres grupos principales.

Tabla 10. Rendimiento de la regresión logística binaria 1-vs-resto (validación cruzada 5-fold)

Grupo	N positivos	Balanced Acc. (DE)	AUC (DE)	Nivel azar
Rodilla vs resto	233	0,527 (0,009)	0,594 (0,030)	0,500
Muslo vs resto	110	0,498 (0,003)	0,584 (0,051)	0,500
Pierna baja vs resto	111	0,500 (0,000)	0,542 (0,043)	0,500

Nota. DE: desviación estándar entre pliegues. N = 755 sujetos en la muestra de modelado.

El modelo de rodilla frente al resto fue el de mejor rendimiento (balanced accuracy = 0,527; AUC = 0,594). Los odds ratios más destacados fueron la velocidad de carrera (OR = 0,532, es decir, una mayor velocidad de carrera se asocia con menor probabilidad de pertenecer al grupo de lesión de rodilla) y la longitud de zancada derecha (OR = 1,586, mayor zancada asociada a mayor probabilidad de rodilla). Para el modelo de muslo, la velocidad presentó el OR más elevado (OR = 2,222), indicando que los corredores más rápidos tienen mayor probabilidad de lesión de muslo, algo coherente con la mayor demanda sobre la musculatura isquiotibial a velocidades altas.

5.7. Modelado predictivo

Los cuatro modelos evaluados mostraron una capacidad discriminativa limitada y bastante homogénea en los tres grupos binarios. Ningún modelo de aprendizaje automático mejoró de forma sustancial sobre la regresión logística.

Tabla 11. Comparación completa de modelos predictivos (validación cruzada 5-fold, N = 755)

Modelo	Grupo	BA	DE(BA)	F1	DE(F1)	AUC	DE(AUC)	Sensib.	Especif.
Regr. logística	Rodilla	0,550	0,030	0,433	0,033	0,591	0,030	0,562	0,538
Regr. logística	Muslo	0,548	0,070	0,259	0,062	0,589	0,047	0,500	0,597
Regr. logística	Pierna baja	0,527	0,046	0,245	0,038	0,540	0,039	0,494	0,559
Árbol decisión	Rodilla	0,549	0,019	0,420	0,044	0,565	0,041	0,537	0,562
Árbol decisión	Muslo	0,526	0,027	0,206	0,073	0,526	0,032	0,400	0,651
Árbol decisión	Pierna baja	0,482	0,037	0,207	0,041	0,480	0,073	0,457	0,507
Random Forest	Rodilla	0,565	0,026	0,377	0,045	0,604	0,026	0,348	0,782
Random Forest	Muslo	0,516	0,016	0,104	0,062	0,558	0,079	0,073	0,960
Random Forest	Pierna baja	0,509	0,018	0,084	0,051	0,535	0,047	0,054	0,964
XGBoost	Rodilla	0,556	0,038	0,386	0,052	0,605	0,032	0,387	0,726
XGBoost	Muslo	0,541	0,042	0,212	0,076	0,569	0,081	0,200	0,882
XGBoost	Pierna baja	0,501	0,018	0,129	0,043	0,518	0,065	0,117	0,885

Nota. BA: balanced accuracy; DE: desviación estándar entre pliegues; Sensib.: sensibilidad; Especif.: especificidad.

Random forest y XGBoost obtuvieron los AUC más altos para rodilla (0,604 y 0,605), mejorando apenas +0,013 y +0,014 sobre la regresión logística (0,591). Sin embargo, esa pequeña mejora en AUC vino acompañada de una caída importante en sensibilidad: el random forest identificó correctamente solo el 34,8% de los casos reales de lesión de rodilla, frente al 56,2% de la regresión logística. Para muslo y pierna baja, la regresión logística fue directamente el mejor modelo. El árbol de decisión fue el que peor funcionó en general, llegando a un AUC de 0,480 para pierna baja, por debajo del nivel de azar.

La velocidad de carrera fue la variable con mayor importancia en random forest para rodilla y muslo. Para pierna baja predominaron el heel whip medial izquierdo y los ángulos de rodilla en el plano frontal, coherente con el mecanismo del síndrome de estrés tibial medial. El árbol de decisión para rodilla realizó su primera división sobre la velocidad estandarizada (umbral ≈ –0,34, equivalente a ~2,5 m/s), confirmando el papel central de esta variable.

Análisis de sensibilidad por valores ausentes.

La variable años de experiencia corriendo presentó un 25,0% de valores ausentes, lo que redujo la muestra de modelado de 979 a 755 sujetos al aplicar eliminación por lista. Para evaluar si esta reducción afecta a los resultados se compararon los modelos de regresión logística en dos configuraciones: con los 11 predictores originales (N = 755) y excluyendo años corriendo para disponer de la muestra completa (N = 979). Los resultados se recogen en la Tabla 12.

Tabla 12. Análisis de sensibilidad: rendimiento de la regresión logística con y sin la variable años de experiencia corriendo

Grupo	N con años corriendo	AUC con años corriendo	N sin años corriendo	AUC sin años corriendo	Diferencia en AUC
Rodilla vs resto	755	0,591	978	0,602	0,011
Muslo vs resto	755	0,589	978	0,608	0,019
Pierna baja vs resto	755	0,540	978	0,564	0,024

Nota. Validación cruzada estratificada de 5 pliegues en ambas configuraciones. La configuración sin años corriendo utiliza los 10 predictores restantes sobre la muestra completa de lesionados.

La exclusión de la variable amplió la muestra en 223 sujetos y produjo mejoras marginales en AUC para rodilla (+0,011) y muslo (+0,019), mientras que pierna baja apenas varió (+0,024). Las mejoras en AUC son consistentes en los tres grupos, pero de magnitud pequeña (+0,011 a +0,024), sin relevancia suficiente para modificar las conclusiones principales: en ambas configuraciones el rendimiento se mantiene en un rango modesto. La variable se mantuvo en el modelo principal porque su aportación está respaldada por la literatura sobre factores de riesgo en corredores.

Figure 6. Model comparison

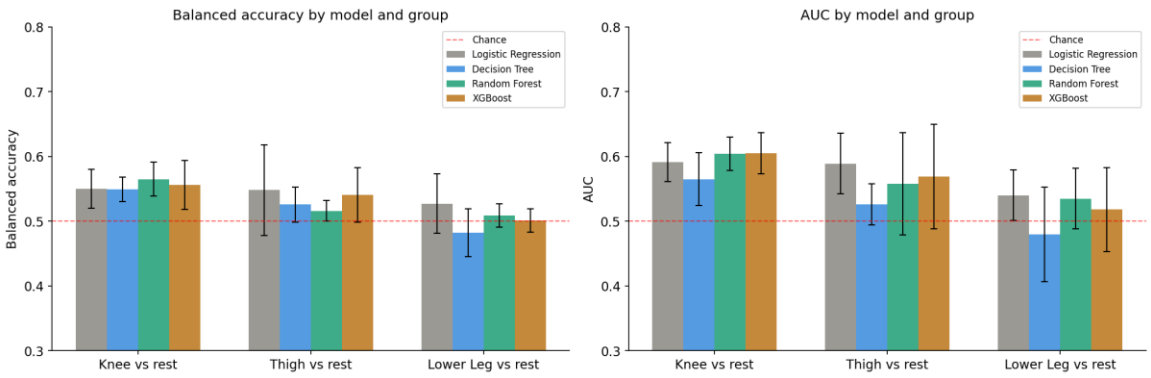


Figura 6. Comparación del rendimiento predictivo. Línea roja: nivel de azar (balanced accuracy = 0,5). Barras de error: DE entre pliegues.

Figure 7. Random Forest feature importance by injury group

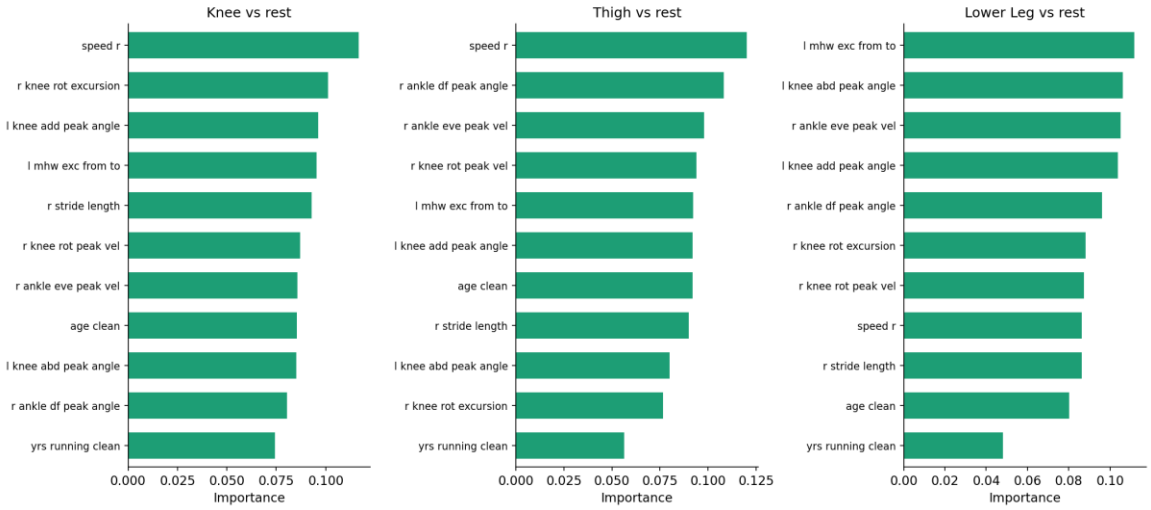


Figura 7. Importancia de variables (mean decrease in Gini impurity) según el random forest.

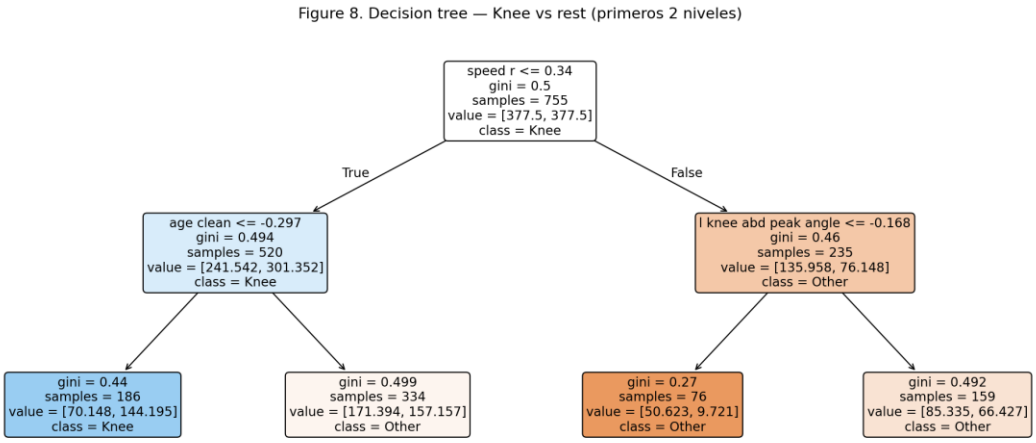


Figura 8. Árbol de decisión para Rodilla vs resto (primeros 2 niveles; max_depth=4, versión completa disponible en el repositorio). Primera división: velocidad de carrera estandarizada.

6. Discusión

Antes de interpretar los resultados, es necesario tener presente una limitación estructural del estudio: el diseño es transversal y retrospectivo. Esto significa que los datos biomecánicos y el historial de lesión se recogieron en un único punto temporal, sin seguimiento longitudinal. Por tanto, los análisis permiten identificar asociaciones entre variables biomecánicas y tipos de lesión, pero no establecer relaciones causales ni determinar si los patrones cinemáticos identificados preceden a la lesión o son consecuencia de ella. Un corredor evaluado con dolor activo o en fase de recuperación muy probablemente ha modificado su patrón de carrera, lo que implica que parte de las diferencias biomecánicas observadas podrían reflejar adaptaciones a la lesión más que factores de riesgo previos. Esta limitación atraviesa toda la interpretación que sigue y debe mantenerse presente al leer los hallazgos.

Los resultados permiten caracterizar los perfiles biomecánicos, antropométricos y deportivos asociados a los principales tipos de lesión en corredores, sobre una muestra de tamaño superior al de la mayoría de estudios previos en este campo. El dataset de Clermont et al. (2024), con 1.402 sujetos únicos, permite detectar efectos con tamaños del efecto tan pequeños como eta² = 0,013 con suficiente potencia estadística.

6.1. Distribución y perfil de la muestra

La distribución de lesiones observada coincide con lo que describen las revisiones sistemáticas de referencia en este campo. La rodilla como localización más frecuente (33,2%), seguida del muslo y la pierna baja, es un patrón consistente con van Gent et al. (2007) y Lopes et al. (2012). El síndrome patelofemoral (12,5%) y el síndrome de la banda iliotibial (10,0%) son también los diagnósticos más citados en esas revisiones para corredores de larga distancia, la muestra reproduce, por tanto, el patrón descrito en la literatura de referencia.

Entre todas las variables demográficas, la única que diferenció significativamente a lesionados de no lesionados fue la velocidad de carrera: los lesionados corrían más despacio. Esta diferencia admite varias lecturas, ninguna excluyente. Correr

más despacio puede ser consecuencia directa de la lesión, una estrategia de protección ante el dolor, o un marcador de una técnica menos eficiente que predispone a lesionarse. El diseño transversal no permite distinguir entre estas opciones. Lo que sí resulta interesante es que el resto de variables demográficas no discriminen entre lesionados y no lesionados como grupo global, pero sí emerjan diferencias claras al comparar los tipos anatómicos entre sí. Esto sugiere que el perfil demográfico no determina si un corredor se lesiona, sino qué tipo de lesión desarrolla.

6.2. Caracterización biomecánica por tipo de lesión

El resultado biomecánico más sólido del análisis es la asociación entre la lesión de tobillo y valores elevados de velocidad y excursión de rotación de rodilla derecha. Los corredores con lesión de tobillo presentaron valores de velocidad de rotación de rodilla notablemente superiores a los de los grupos de rodilla, cadera/pelvis, pierna baja y muslo, con medianas de 224,0 °/s frente a valores entre 184,6 y 198,1 °/s en esos grupos. Desde un punto de vista biomecánico, esto tiene una explicación razonable: una mayor rotación de rodilla en el plano transversal durante la fase de apoyo se asocia con mayor carga torsional transmitida hacia abajo a través de la cadena cinética cerrada, lo que podría incrementar las fuerzas sobre el tendón de Aquiles y los ligamentos del tobillo. Powers (2010) describió exactamente este mecanismo de amplificación distal de alteraciones cinemáticas proximales.

No obstante, la interpretación de estas diferencias biomecánicas requiere una cautela metodológica importante. La velocidad de carrera difiere significativamente entre grupos de lesión (eta² = 0,022) y está fisiológicamente asociada a numerosas variables cinemáticas, entre ellas la longitud de zancada, las velocidades angulares y los tiempos de apoyo. En consecuencia, no es posible determinar con este diseño en qué medida las diferencias biomecánicas observadas entre grupos reflejan patrones cinemáticos propios de cada tipo de lesión, o son al menos parcialmente atribuibles a las diferencias de velocidad entre

los corredores de cada grupo. Esta ambigüedad es inherente al diseño transversal y solo podría resolverse mediante análisis con velocidad controlada o estratificación por grupos de velocidad, lo que queda como línea de trabajo futura.

Los patrones de aducción y abducción de rodilla izquierda también ofrecen información relevante. Los grupos de pierna baja y muslo presentaron valores de aducción de rodilla menos pronunciados que los demás grupos, formando un cluster diferenciado en el post-hoc de Dunn. Este patrón es coherente con lo documentado por Noehren et al. (2007) y Neal et al. (2016), que identificaron la aducción de cadera y la rotación interna de rodilla como factores biomecánicos diferenciadores en corredores que desarrollan síndrome de la banda iliotibial y síndrome patelofemoral respectivamente.

La velocidad pico de eversión de tobillo derecho generó el post-hoc más limpio y simétrico del conjunto: Ankle, Lower Leg y Thigh presentaron mayor velocidad de eversión que Knee y Other, con el mismo valor de p en todos los contrastes significativos. Tobillo, pierna baja y muslo parecen compartir un mecanismo de pronación rápida, que probablemente genera fuerzas de torsión sobre la tibia y sobrecarga el compartimento muscular posterior.

El gradiente de edad y experiencia que emerge de los datos es clínicamente interpretable y bastante intuitivo: los corredores más jóvenes y menos experimentados concentran más lesiones de rodilla y muslo, posiblemente por errores de planificación del entrenamiento o déficits en el control neuromuscular que mejoran con la práctica. Los más veteranos tienden a lesionarse en estructuras sometidas a mayor carga acumulada a lo largo de los años, como el tendón de Aquiles o la articulación coxofemoral.

Es importante recordar, no obstante, que el diseño transversal impide confirmar si estas diferencias de perfil demográfico preceden a la lesión o son el resultado de años de práctica con patrones de movimiento ya condicionados por lesiones previas.

6.3. Capacidad discriminativa de los modelos

Todos los modelos predictivos mostraron una capacidad discriminativa limitada y homogénea, con AUC entre 0,48 y 0,61. Este resultado es coherente con lo reportado en revisiones sobre machine learning en lesiones deportivas (Van Eetvelde et al., 2021) y con las conclusiones de Peterson et al. (2022), que encontró que las variables biomecánicas disponibles en la literatura tienen tamaños del efecto triviales a pequeños cuando se usan como factores de riesgo independientes en corredores no élite.

El hecho de que ningún modelo de mayor complejidad supere a la regresión logística constituye en sí mismo un resultado relevante. Apunta a que las relaciones entre variables y tipo de lesión son principalmente aditivas y que las variables disponibles no aportan suficiente información discriminativa, independientemente del nivel de complejidad del modelo.

Resulta también relevante señalar que la velocidad de carrera fue la variable con mayor importancia en los modelos predictivos para rodilla y muslo, y la primera división del árbol de decisión. Esto refuerza su papel central en el análisis, pero también su potencial como factor de confusión: un modelo que depende principalmente de la velocidad para clasificar el tipo de lesión puede estar capturando en parte diferencias sistemáticas en el perfil del corredor más que patrones biomecánicos específicos de cada lesión.

La regresión logística ofreció el mejor equilibrio entre sensibilidad y especificidad y produce odds ratios interpretables clínicamente, lo que la convierte en el modelo más adecuado para aplicaciones de cribado donde identificar correctamente los casos positivos es prioritario.

6.4. Implicaciones clínicas

Los resultados de este trabajo no permiten construir un clasificador clínicamente útil para predecir el tipo de lesión de un corredor a partir de su perfil biomecánico. Sin embargo, sí ofrecen patrones concretos que pueden orientar decisiones clínicas en contextos específicos.

El hallazgo más robusto, la asociación entre mayor velocidad y excursión de rotación de rodilla derecha y lesión de tobillo, apunta a una hipótesis de intervención que convendría explorar en estudios prospectivos: si la reducción de esa rotación mediante correcciones posturales o aumento de cadencia tuviera efecto sobre la sintomatología, podría constituir una línea de reentrenamiento. Heiderscheit et al. (2011) demostraron que aumentar la cadencia un 5–10% modifica la cinemática de rodilla en corredores; si ese efecto alcanzara también a la rotación en el plano transversal, podría constituir una diana de reentrenamiento, aunque esa transferencia no ha sido evaluada directamente.

El patrón simétrico de la velocidad de eversión de tobillo derecho, que diferenció a los grupos de tobillo, pierna baja y muslo del resto, apunta a un mecanismo común de pronación rápida en esas tres categorías. En la práctica clínica, este patrón apuntaría a que una valoración conjunta de la cadena posterior en corredores con cualquiera de esas tres lesiones podría ser informativa, aunque los datos disponibles no permiten evaluar el efecto de esa estrategia sobre los resultados clínicos.

El gradiente de edad y experiencia tiene implicaciones para la prevención primaria. Los corredores más jóvenes y con menos experiencia concentran las lesiones de rodilla y muslo, mientras que los más veteranos acumulan más lesiones en tobillo, pie y cadera. Este patrón, si se confirmara en estudios prospectivos, apuntaría a que las estrategias preventivas podrían diferir según el perfil del corredor: en los más jóvenes, la planificación del entrenamiento y el aumento gradual del volumen; en los más experimentados, la salud tendinosa y articular a largo plazo.

6.5. Implicaciones metodológicas

Este trabajo pone de manifiesto un aspecto que en biomecánica no siempre recibe suficiente atención: trabajar con muestras grandes permite detectar diferencias estadísticamente significativas con tamaños del efecto muy pequeños, pero eso no equivale automáticamente a relevancia clínica. Con 1.402 sujetos fue posible detectar diferencias con eta² tan pequeño como 0,013, pero todos los efectos encontrados tienen un solapamiento sustancial entre grupos que los hace poco útiles para clasificación individual.

La selección de predictores realizada antes de la validación cruzada puede introducir un optimismo moderado en las estimaciones de rendimiento, algo que queda documentado explícitamente como limitación. Dado que los AUC obtenidos son bajos en todos los casos, este sesgo no cambia las conclusiones principales.

7. Limitaciones

La limitación principal es el diseño transversal. Con este tipo de datos no hay forma de saber si los patrones biomecánicos encontrados existían antes de la lesión o si son consecuencia de ella. Un corredor que lleva semanas con dolor en el tobillo muy probablemente ha modificado su patrón de carrera para protegerse, lo que significa que las diferencias biomecánicas observadas podrían reflejar adaptaciones a la lesión más que factores de riesgo previos. Sin un diseño prospectivo que evalúe a los corredores antes de que se lesionen, esta pregunta no tiene respuesta con estos datos.

Los datos de lesión se obtuvieron por autoinforme, con dos fuentes de error bien documentadas en la literatura. La primera es el sesgo de recuerdo: los corredores no siempre recuerdan con precisión cuándo ocurrió la lesión ni cómo evolucionó. La segunda es el sesgo de clasificación: el campo de diagnóstico específico tuvo un 25,9% de valores ausentes y una variabilidad terminológica notable, con nombres distintos para lo que probablemente es la misma entidad clínica.

La decisión de conservar un único registro por sujeto fue metodológicamente necesaria para evitar la pseudorreplicación, pero tuvo un coste: se descartaron 430 registros adicionales que en algunos casos recogían cambios en el estado de lesión entre sesiones. Esa información longitudinal podría haber sido muy valiosa para estudiar la evolución de los patrones biomecánicos en relación con la lesión.

Setenta y dos variables biomecánicas presentaron varianza nula en la versión exportada del dataset y tuvieron que excluirse. Su recalculación desde los datos crudos podría revelar información adicional relevante, especialmente en dominios como la pelvis o la pronación, donde el número de variables disponibles quedó muy reducido.

La selección de predictores se realizó con el dataset completo antes de la validación cruzada, lo que introduce un optimismo moderado en las estimaciones de rendimiento. Como se ha mencionado en la discusión, dado que los AUC son bajos en todos los casos, este sesgo no altera las conclusiones principales, pero queda documentado aquí de forma explícita.

La muestra está enriquecida con casos lesionados al proceder principalmente de clínicas y centros especializados. La prevalencia observada del 69,8% no es representativa de la población general de corredores y no debe interpretarse como tal.

La ausencia de variables de carga de entrenamiento es una limitación relevante que va más allá de lo metodológico. Los kilómetros semanales, la variabilidad de carga o el historial de entrenamiento reciente son factores de riesgo reconocidos para lesiones por uso excesivo, y su inclusión podría haber mejorado tanto la caracterización de los grupos como el rendimiento predictivo de los modelos. Simplemente no estaban disponibles en este dataset.

8. Objetivos de Desarrollo Sostenible

Este trabajo se vincula principalmente con tres Objetivos de Desarrollo Sostenible de la Agenda 2030 de Naciones Unidas.

ODS 3 Salud y bienestar.

El objetivo central del trabajo es mejorar la comprensión de los factores asociados a las lesiones musculoesqueléticas en corredores, una de las causas más frecuentes de reducción de la actividad física en la población adulta.
Identificar qué perfiles biomecánicos y deportivos se asocian con distintos tipos de lesión contribuye directamente a la prevención y a la promoción de una práctica deportiva más saludable.
La actividad física regular es uno de los pilares de la salud pública recogidos en este ODS, y reducir la incidencia de lesiones es una condición necesaria para sostenerla a largo plazo.

ODS 10 Reducción de las desigualdades.

La carrera a pie es uno de los deportes más accesibles económicamente, practicado por personas de muy distintos niveles socioeconómicos.
Las herramientas de análisis biomecánico han estado históricamente disponibles solo en contextos de élite o en clínicas privadas de alto coste.
Este trabajo utiliza un dataset público y desarrolla metodología reproducible en entornos de código abierto, contribuyendo a democratizar el acceso a análisis biomecánicos de calidad.
Todo el código está disponible en abierto en el repositorio de GitHub asociado al trabajo.

ODS 9 Industria, innovación e infraestructura.

La integración de técnicas de ciencia de datos y aprendizaje automático con datos biomecánicos clínicos representa una línea de innovación metodológica con potencial de aplicación en sistemas de cribado deportivo y herramientas de apoyo a la decisión clínica.

Aunque los resultados muestran las limitaciones actuales de estos modelos con las variables disponibles, el análisis está completamente documentado y es reproducible, lo que lo convierte en un punto de partida concreto para estudios futuros con mejores datos.

9. Conclusiones y líneas futuras

Este trabajo abordó una pregunta: ¿es posible identificar qué variables biomecánicas, antropométricas y deportivas se asocian con cada tipo de lesión en corredores, y hasta qué punto permiten predecirla? La respuesta, después de analizar 1.402 sujetos y evaluar 80 variables biomecánicas con cuatro modelos distintos, es matizada pero clara.

Los datos muestran diferencias reales entre grupos de lesión; varias de ellas tienen una lectura clínica razonable. Los corredores con distintos tipos de lesión difieren significativamente en edad, velocidad de carrera y años de experiencia, y el patrón que emerge tiene sentido: los más jóvenes y menos experimentados concentran más lesiones de rodilla y muslo, mientras que los de mayor edad y experiencia acumulan más lesiones en tobillo, pie y cadera. Esto apunta a mecanismos distintos según el perfil del corredor, posiblemente relacionados con déficits de control neuromuscular en los más jóvenes y con la carga acumulada sobre estructuras tendinosas y articulares en los más veteranos. Estas interpretaciones son, no obstante, necesariamente tentativas: el diseño transversal no permite establecer si los perfiles identificados anteceden a la lesión o la reflejan.

A nivel biomecánico, nueve de las 80 variables testadas mostraron diferencias significativas tras corregir por comparaciones múltiples. El hallazgo más robusto es la asociación entre la velocidad y la excursión de rotación de rodilla derecha y la lesión de tobillo: los corredores de ese grupo presentaron valores significativamente más altos que los de rodilla, cadera, pierna baja y muslo, un patrón coherente con los mecanismos de transmisión de carga torsional a través de la cadena cinética cerrada descritos en la literatura. La velocidad pico de eversión de tobillo derecho mostró el post-hoc más limpio y simétrico del conjunto, diferenciando a los grupos de tobillo, pierna baja y muslo de los de rodilla y otras localizaciones, lo que sugiere un mecanismo común de pronación rápida en esas tres categorías. El cluster formado por pierna baja y muslo, con menor aducción de rodilla izquierda que el resto de grupos, es también un resultado consistente e interpretable desde la biomecánica.

Ahora bien, la magnitud de estos efectos es pequeña. Ningún eta² superó 0,03, lo que significa que el solapamiento entre grupos es muy grande. Hay diferencias estadísticamente significativas, pero el solapamiento entre grupos es tan sustancial que no permiten clasificar a un corredor individual en un tipo de lesión a partir de su perfil biomecánico. Este resultado no refleja una limitación del análisis, sino una conclusión sustantiva, y es coherente con lo que reportan las revisiones sistemáticas más recientes sobre biomecánica de la carrera y lesiones.

En cuanto a los modelos predictivos, el rendimiento fue limitado y homogéneo en los tres grupos binarios evaluados, con AUC entre 0,48 y 0,61. El modelo multinomial con los siete grupos simultáneos no superó el nivel de azar, lo que llevó a adoptar la estrategia de modelos binarios 1-vs-resto. Los modelos de aprendizaje automático, random forest y XGBoost, obtuvieron AUC marginalmente superiores a la regresión logística para el grupo de rodilla, pero esa pequeña mejora vino acompañada de una caída importante en sensibilidad que los hace poco adecuados para aplicaciones de cribado clínico, donde identificar correctamente los casos positivos es prioritario. Para muslo y pierna baja, la regresión logística fue directamente el mejor modelo. El hecho de que modelos más complejos no mejoren sobre un modelo lineal sugiere que las relaciones entre las variables disponibles y el tipo de lesión son principalmente aditivas, y que el problema no es la complejidad del modelo sino la cantidad de información discriminativa que contienen las variables.

Desde el punto de vista metodológico, este trabajo pone de manifiesto un aspecto relevante: trabajar con muestras grandes en biomecánica permite detectar diferencias con tamaños del efecto muy pequeños, y eso es valioso porque da una imagen más realista de lo que la biomecánica puede y no puede explicar sobre las lesiones en corredores. Pero también obliga a distinguir con cuidado entre significación estadística y relevancia clínica, una distinción que en la literatura de este campo no siempre se hace con suficiente rigor.

La biomecánica permite caracterizar el perfil de lesión de un corredor, pero las variables disponibles en este dataset no tienen suficiente poder discriminativo para construir herramientas de clasificación con utilidad clínica real. Lo que sí ofrece son patrones concretos que pueden orientar intervenciones específicas, como el reentrenamiento de la cinemática de rodilla en corredores con tendinopatía aquílea, o los programas de prevención centrados en la planificación del entrenamiento para corredores jóvenes con poca experiencia.

Como consideración metodológica que atraviesa todas las conclusiones anteriores: el diseño transversal y retrospectivo impide inferir causalidad. Los resultados describen asociaciones observadas en un único corte temporal y no permiten afirmar que ninguna de las variables identificadas cause, predisponga o proteja frente a un tipo concreto de lesión. Esta distinción entre asociación y causalidad es especialmente relevante en un campo como la biomecánica de la carrera, donde la tentación de interpretar diferencias cinemáticas como factores de riesgo modificables es habitual en la literatura.

Por tamaño muestral, este trabajo se sitúa entre los análisis biomecánicos de lesiones en corredores más amplios realizados hasta la fecha y establece una base metodológica sólida para futuros estudios prospectivos que incorporen variables de carga de entrenamiento y que evalúen a los corredores antes de que se lesionen. Esos estudios son los que, en última instancia, podrán responder las preguntas que este diseño transversal deja abiertas.

Líneas futuras de investigación

Los resultados de este trabajo apuntan a varias direcciones prioritarias para investigación futura. La más relevante es el diseño prospectivo: evaluar biomecánicamente a corredores antes de que se lesionen y hacer seguimiento longitudinal permitiría determinar si los patrones identificados aquí preceden a la lesión o son consecuencia de ella, la pregunta que este diseño transversal deja abierta.

La incorporación de variables de carga de entrenamiento (kilómetros semanales, variabilidad de carga, historial de entrenamiento reciente) mejoraría sustancialmente tanto la caracterización de los grupos como el rendimiento predictivo de los modelos. Estas variables son factores de riesgo reconocidos que simplemente no estaban disponibles en este dataset.

El recalculo de las 72 variables biomecánicas con varianza nula a partir de los datos crudos del sistema de captura ampliaría el espacio de predictores disponibles, especialmente en dominios como la pelvis y la pronación donde el número de variables quedó muy reducido.

Finalmente, la aplicación de modelos más sofisticados de aprendizaje automático, como redes neuronales sobre series temporales de la señal cinemática completa en lugar de variables resumen, podría capturar patrones que los descriptores escalares no recogen. Esa línea requeriría datos en formato de series temporales, no disponibles en la versión pública del dataset utilizado.

Todo el código empleado en el análisis está disponible públicamente en https://github.com/TheMatrix5/tfg-running-lesiones, organizado por fases y documentado para facilitar su reproducción.

La estructura completa del repositorio de código se recoge en la Tabla A12 del anexo.

10. Anexos

Las siguientes tablas complementan los análisis presentados en el cuerpo del trabajo. Se incluyen aquí por razones de extensión y para no interrumpir el flujo argumentativo de los resultados.

Tabla A1. Variables antropométricas, deportivas y de lesión

Variable	Tipo	% ausentes	Decisión	Motivo
speed_r	float	0,0%	KEEP	Velocidad de carrera — predictor clave
age	int	0,0%	KEEP	Edad — covariable demográfica
Height	float	0,1%	KEEP	Talla — variable antropométrica
Weight	float	0,1%	KEEP	Peso — variable antropométrica
Gender	texto	0,0%	KEEP	Género — covariable demográfica
DominantLeg	texto	14,7%	KEEP	Pierna dominante — lateralidad
Level	texto	8,8%	KEEP	Nivel recreativo/competitivo
YrsRunning	float	25,0%	KEEP*	Experiencia — usar versión limpia
Activities	texto	11,9%	DROP	Alta variabilidad, baja relevancia
InjDuration	float	83,0%	DROP	Missings > 50%
NumRaces	float	78,9%	DROP	Missings > 50%
YrPR	float	84,5%	DROP	Missings > 50%
InjDefn	texto	5,3%	OUTCOME	Severidad — variable respuesta
InjJoint	texto	16,0%	OUTCOME	Localización — base de la variable objetivo
InjSide	texto	18,8%	OUTCOME	Lado afectado
SpecInjury	texto	25,9%	OUTCOME	Diagnóstico — descriptivo
SpecInjury2	texto	79,5%	DROP	Missings > 50%

Nota. KEEP: incluida. KEEP: incluida con cautela. DROP: excluida. OUTCOME: usada para construir la variable dependiente.*

Tabla A2. Variables biomecánicas — resumen por dominio

Dominio anatómico	Total	KEEP	DROP (var=0)	Motivo
Rodilla (izq + der)	40	20	20	Varianza nula — no calculables
Cadera (izq + der)	32	16	16	Varianza nula — no calculables
Tobillo (izq + der)	32	16	16	Varianza nula — no calculables
Pie (izq + der)	10	8	2	Varianza nula — no calculables
Pelvis (izq + der)	8	4	4	Varianza nula — no calculables
Pronación (izq + der)	8	4	4	Varianza nula — no calculables
Temporal-espaciales	22	12	10	Varianza nula — no calculables
TOTAL	152	80	72	—

Nota. Las 72 variables excluidas no pudieron calcularse en el proceso de exportación desde el software de análisis biomecánico original.

Tabla A4. Dunn: LEFT_STRIDE_LENGTH

Patrón idéntico a RIGHT_STRIDE_LENGTH (r = 1,000). Other presenta las zancadas más cortas, diferenciándose de Ankle, Knee, Lower Leg y Thigh.

	Ankle	Foot	Hip/Pelvis	Knee	Lower Leg	Other	Thigh
Ankle	—	**0,0082****	0,0975	0,1519	0,4892	**0,0010****	0,7345
Foot	**0,0082****	—	0,3302	0,0816	**0,0272***	0,3102	**0,0074****
Hip/Pelvis	0,0975	0,3302	—	0,5243	0,2413	0,0654	0,1046
Knee	0,1519	0,0816	0,5243	—	0,4002	**0,0072****	0,1629
Lower Leg	0,4892	**0,0272***	0,2413	0,4002	—	**0,0025****	0,6120
Other	**0,0010****	0,3102	0,0654	**0,0072****	**0,0025****	—	**0,0010****
Thigh	0,7345	**0,0074****	0,1046	0,1629	0,6120	**0,0010****	—

*Nota Nota. La correlación bilateral perfecta justificó excluir LEFT_STRIDE_LENGTH del modelado. *** p < 0,001; ** p < 0,01; * p < 0,05; sin color: ns (p ≥ 0,05). Todos los valores p corresponden al test post-hoc de Dunn con corrección FDR.*

Tabla A5. Dunn: LEFT_KNEE_ADD_PEAK_ANGLE

Lower Leg y Thigh forman un cluster diferenciado del resto con menor aducción de rodilla. Se diferencian de Ankle (p = 0,041), Foot (p = 0,009), Hip/Pelvis (p = 0,019) y Knee (p = 0,019), pero no entre sí (p = 0,992).

La aducción de rodilla izquierda muestra un patrón especialmente llamativo: los grupos de pierna baja y muslo forman un bloque diferenciado del resto, con valores de aducción menos pronunciados que los demás grupos.

	Ankle	Foot	Hip/Pelvis	Knee	Lower Leg	Other	Thigh
Ankle	—	0,8549	0,9541	0,9541	**0,0407***	0,9541	**0,0407***
Foot	0,8549	—	0,8549	0,4444	**0,0090****	0,8549	**0,0090****
Hip/Pelvis	0,9541	0,8549	—	0,8549	**0,0187***	0,9923	**0,0187***
Knee	0,9541	0,4444	0,8549	—	**0,0187***	0,8549	**0,0187***
Lower Leg	**0,0407***	**0,0090****	**0,0187***	**0,0187***	—	**0,0407***	0,9923
Other	0,9541	0,8549	0,9923	0,8549	**0,0407***	—	**0,0407***
Thigh	**0,0407***	**0,0090****	**0,0187***	**0,0187***	0,9923	**0,0407***	—

*Nota. Patrón en bloque: Lower Leg + Thigh vs el resto es el más llamativo de todo el conjunto post-hoc. *** p < 0,001; ** p < 0,01; * p < 0,05; sin color: ns (p ≥ 0,05). Todos los valores p corresponden al test post-hoc de Dunn con corrección FDR.*

Tabla A6. Dunn: LEFT_KNEE_ABD_PEAK_ANGLE

Lower Leg se diferencia de Foot (p = 0,010) e Hip/Pelvis (p = 0,025). Foot también difiere de Thigh (p = 0,025). Patrón más disperso que en aducción.

La abducción de rodilla izquierda sigue un patrón más disperso que la aducción, con diferencias más marginales entre grupos, coherente con su menor tamaño del efecto (eta² = 0,015).

	Ankle	Foot	Hip/Pelvis	Knee	Lower Leg	Other	Thigh
Ankle	—	0,3603	0,6592	0,6592	0,1230	0,6640	0,2395
Foot	0,3603	—	0,6640	0,0816	**0,0099****	0,6592	**0,0245***
Hip/Pelvis	0,6592	0,6640	—	0,2315	**0,0245***	0,8886	0,0699
Knee	0,6592	0,0816	0,2315	—	0,1230	0,3603	0,3106
Lower Leg	0,1230	**0,0099****	**0,0245***	0,1230	—	0,0699	0,6640
Other	0,6640	0,6592	0,8886	0,3603	0,0699	—	0,1230
Thigh	0,2395	**0,0245***	0,0699	0,3106	0,6640	0,1230	—

*Nota. Las diferencias son más marginales, coherente con el menor eta² de esta variable (0,015). *** p < 0,001; ** p < 0,01; * p < 0,05; sin color: ns (p ≥ 0,05). Todos los valores p corresponden al test post-hoc de Dunn con corrección FDR.*

Tabla A7. Dunn: RIGHT_ANKLE_EVE_PEAK_VEL

Patrón muy simétrico: Ankle, Lower Leg y Thigh tienen mayor velocidad de eversión que Knee y Other, con p = 0,033 en todos los contrastes significativos.

La velocidad pico de eversión de tobillo derecho genera el post-hoc más limpio y simétrico del conjunto: los grupos de tobillo, pierna baja y muslo presentan mayor velocidad de eversión que rodilla y otras localizaciones, con el mismo valor de p en todos los contrastes significativos.

	Ankle	Foot	Hip/Pelvis	Knee	Lower Leg	Other	Thigh
Ankle	—	0,5460	0,5055	**0,0327***	0,8350	**0,0327***	0,8350
Foot	0,5460	—	0,8904	0,2083	0,5982	0,1761	0,5982
Hip/Pelvis	0,5055	0,8904	—	0,2941	0,5460	0,2083	0,5460
Knee	**0,0327***	0,2083	0,2941	—	**0,0327***	0,5982	**0,0327***
Lower Leg	0,8350	0,5982	0,5460	**0,0327***	—	**0,0327***	0,9709
Other	**0,0327***	0,1761	0,2083	0,5982	**0,0327***	—	**0,0327***
Thigh	0,8350	0,5982	0,5460	**0,0327***	0,9709	**0,0327***	—

*Nota. El post-hoc más limpio y simétrico del conjunto. *** p < 0,001; ** p < 0,01; * p < 0,05; sin color: ns (p ≥ 0,05). Todos los valores p corresponden al test post-hoc de Dunn con corrección FDR.*

Tabla A8. Dunn: RIGHT_ANKLE_DF_PEAK_ANGLE

Solo un contraste significativo: Hip/Pelvis vs Knee (p = 0,049). El efecto está distribuido de forma difusa entre los pares.

El ángulo pico de dorsiflexión de tobillo derecho mostró significación global en el test de Kruskal-Wallis, pero el post-hoc revela que solo hay un contraste significativo: cadera/pelvis frente a rodilla. El efecto está distribuido de forma difusa entre los pares.

	Ankle	Foot	Hip/Pelvis	Knee	Lower Leg	Other	Thigh
Ankle	—	0,9229	0,9229	0,0714	0,1052	0,6759	0,1128
Foot	0,9229	—	0,9229	0,0714	0,1029	0,6759	0,1101
Hip/Pelvis	0,9229	0,9229	—	**0,0488***	0,0714	0,5643	0,0846
Knee	0,0714	0,0714	**0,0488***	—	0,9229	0,3010	0,8941
Lower Leg	0,1052	0,1029	0,0714	0,9229	—	0,4633	0,9229

	Ankle	Foot	Hip/Pelvis	Knee	Lower Leg	Other	Thigh
Other	0,6759	0,6759	0,5643	0,3010	0,4633	—	0,5226
Thigh	0,1128	0,1101	0,0846	0,8941	0,9229	0,5226	—

*Nota. Variable con significación global en KW pero con pocos contrastes post-hoc significativos. *** p < 0,001; ** p < 0,01; * p < 0,05; sin color: ns (p ≥ 0,05). Todos los valores p corresponden al test post-hoc de Dunn con corrección FDR.*

Tabla A9. Dunn: LEFT_MHW_EXC_FROM_TO

Lower Leg tiene el menor heel whip medial y se diferencia de Ankle (p = 0,040) e Hip/Pelvis (p = 0,040). Ankle e Hip/Pelvis tienen la mayor excursión.

La excursión del heel whip medial izquierdo diferencia principalmente al grupo de pierna baja, que presenta la menor excursión y se distingue significativamente de tobillo y cadera/pelvis. Esta variable resultó ser la de mayor importancia en el random forest para el grupo de pierna baja.

	Ankle	Foot	Hip/Pelvis	Knee	Lower Leg	Other	Thigh
Ankle	—	0,7735	0,9085	0,1311	**0,0401***	0,0945	0,7347
Foot	0,7735	—	0,8015	0,2013	0,0513	0,1311	0,9085
Hip/Pelvis	0,9085	0,8015	—	0,1229	**0,0401***	0,0945	0,7532
Knee	0,1311	0,2013	0,1229	—	0,2013	0,5578	0,2013
Lower Leg	**0,0401***	0,0513	**0,0401***	0,2013	—	0,8560	0,0513
Other	0,0945	0,1311	0,0945	0,5578	0,8560	—	0,1311
Thigh	0,7347	0,9085	0,7532	0,2013	0,0513	0,1311	—

*Nota. El heel whip medial fue la variable de mayor importancia en el random forest para pierna baja. *** p < 0,001; ** p < 0,01; * p < 0,05; sin color: ns (p ≥ 0,05). Todos los valores p corresponden al test post-hoc de Dunn con corrección FDR.*

En conjunto, el análisis post-hoc confirma que el grupo Ankle es el más diferenciado del resto, especialmente en las variables de rotación de rodilla. El cluster formado por Lower Leg y Thigh, con menor aducción de rodilla izquierda, y el patrón simétrico de la velocidad de eversión de tobillo derecho son los hallazgos más consistentes entre el resto de variables.

Tabla A10. Versiones de librerías del entorno de análisis

Librería	Versión	Uso
Python	3.9.13	Lenguaje base
pandas	2.2.3	Manipulación de datos
numpy	1.24.4	Operaciones numéricas
scipy	1.13.1	Tests estadísticos
scikit-learn	≥ 1.0	Modelos ML y validación cruzada
statsmodels	≥ 0.14	Corrección FDR
scikit-posthocs	≥ 0.7	Test post-hoc de Dunn
xgboost	≥ 2.0	Modelo XGBoost
matplotlib	3.9.2	Figuras
seaborn	0.13.2	Visualizaciones estadísticas

Nota. Semilla aleatoria: random_state = 42. SO: Windows 10.

Tabla A11. Hiperparámetros de los modelos predictivos

Modelo	Parámetros	Notas
Regr. logística	solver='lbfgs', C=1.0, max_iter=1000, class_weight='balanced', random_state=42	Regularización L2. class_weight='balanced' compensa el desbalance.
Árbol de decisión	max_depth=4, class_weight='balanced', criterion='gini', random_state=42	Profundidad limitada para interpretabilidad.
Random Forest	n_estimators=200, max_depth=6, class_weight='balanced_subsample', n_jobs=-1, random_state=42	balanced_subsample maneja el desbalance por submuestreo bootstrap.
XGBoost	n_estimators=200, max_depth=4, learning_rate=0.05, scale_pos_weight=neg/pos, eval_metric='logloss', random_state=42	scale_pos_weight calculado por grupo binario como N_neg/N_pos.

Nota. StratifiedKFold(n_splits=5, shuffle=True, random_state=42). Estandarización (StandardScaler) dentro del pipeline en cada pliegue para evitar data leakage. Sin búsqueda de hiperparámetros.

Tabla A12. Estructura de notebooks del repositorio

Notebook	Contenido
01_fase1_preparacion.ipynb	Carga, auditoría, tabla maestra de variables, variable objetivo, versiones limpias
02_fase2_descriptivo.ipynb	Descriptivo completo, Shapiro-Wilk, distribución de lesiones
03_fase3_caracterizacion.ipynb	Kruskal-Wallis (80 vars), FDR, Dunn post-hoc, heatmap, boxplots
04_fase4_modelado_explicativo.ipynb	Regresión logística multinomial y binaria, OR con IC 95% bootstrap
05_fase5_modelado_predictivo.ipynb	Árbol, Random Forest, XGBoost, comparación de modelos, feature importance

10. Referencias bibliográficas

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological), 57*(1), 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Chen, T., & 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

Clermont, C. A., Phinyomark, A., Lusina, S. J., & Ferber, R. (2024). A dataset of running biomechanics and injury history from recreational to competitive runners. *Scientific Data, 11*, 32. https://doi.org/10.1038/s41597-024-04011-7

Peterson, B., Hawke, F., Spink, M., Sadler, S., Hawes, M., Callister, R., & Chuter, V. (2022). Biomechanical and musculoskeletal measurements as risk factors for running-related injury in non-elite runners: a systematic review and meta-analysis of prospective studies. *Sports Medicine – Open, 8*(1), 38. https://doi.org/10.1186/s40798-022-00416-z

Dunn, O. J. (1964). Multiple comparisons using rank sums. *Technometrics, 6*(3), 241–252. https://doi.org/10.1080/00401706.1964.10490181

Heiderscheit, B. C., Chumanov, E. S., Michalski, M. P., Wille, C. M., & Ryan, M. B. (2011). Effects of step rate manipulation on joint mechanics during running. *Medicine & Science in Sports & Exercise, 43*(2), 296–302. https://doi.org/10.1249/MSS.0b013e3181ebedf4

Lopes, A. D., Hespanhol Júnior, L. C., Yeung, S. S., & Costa, L. O. P. (2012). What are the main running-related musculoskeletal injuries? A systematic review. *Sports Medicine, 42*(10), 891–905. https://doi.org/10.1007/BF03262301

Neal, B. S., Lack, S. D., Lankhorst, N. E., Raye, A., Morrissey, D., & van Middelkoop, M. (2016). Risk factors for patellofemoral pain: a systematic review and meta-analysis. *British Journal of Sports Medicine, 50*(14), 844–852. https://doi.org/10.1136/bjsports-2015-094904

Noehren, B., Davis, I., & Hamill, J. (2007). Prospective study of the biomechanical factors associated with iliotibial band syndrome. *Clinical Biomechanics, 22*(9), 951–956. https://doi.org/10.1016/j.clinbiomech.2007.07.001

Powers, C. M. (2010). The influence of abnormal hip mechanics on knee injury: a biomechanical perspective. *Journal of Orthopaedic & Sports Physical Therapy, 40*(2), 42–51. https://doi.org/10.2519/jospt.2010.3337

Van der Worp, M. P., ten Haaf, D. S. M., van Cingel, R., de Wijer, A., Nijhuis-van der Sanden, M. W. G., & Staal, J. B. (2016). Injuries in runners: a systematic review on risk factors and sex differences. *PLOS ONE, 11*(2), e0147567. https://doi.org/10.1371/journal.pone.0147567

Van Eetvelde, H., Mendonça, L. D., Ley, C., Seil, R., & 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

van Gent, R. N., Siem, D., van Middelkoop, M., van Os, A. G., Bierma-Zeinstra, S. M. A., & Koes, B. W. (2007). Incidence and determinants of lower extremity running injuries in long distance runners: a systematic review. *British Journal of Sports Medicine, 41*(8), 469–480. https://doi.org/10.1136/bjsm.2006.033548

Universidad Europea VALENCIA

ANEXO - DECLARACIÓN DEL USO DE INTELIGENCIA ARTIFICIAL

En la elaboración del presente TFG se han utilizado herramientas de Inteligencia Artificial (IA) como apoyo en distintas fases del trabajo académico.

Herramientas utilizadas: Claude

Estas herramientas se han utilizado para las siguientes tareas:

- Apoyo a la redacción y mejora de estilo académico.
- Generación, revisión o depuración de código software.

DECLARO QUE:

- El contenido final presentado es resultado de mi trabajo personal y académico.
- He revisado y validado críticamente toda la información obtenida mediante herramientas de IA.
- El uso de IA no ha sustituido el análisis, la reflexión ni las conclusiones propias exigidas en el TFG.
- El trabajo cumple las normas de integridad académica y propiedad intelectual establecidas por la universidad.

Nombre del estudiante: Matías Perez Lasala Ciacciarelli

Firma:



Escaneado con CamScanner