GRADO EN INGENIERÍA EN TECNOLOGÍAS DE TELECOMUNICACIÓN

TRABAJO FIN DE GRADO

Estudio de variables relevantes y análisis predictivo en la probabilidad de lesión de futbolistas en los campeonatos de Latinoamérica

Autor: María Ángeles Moreno Nieto

Director: Luis Francisco Sánchez Merchante

Madrid

Declaro, bajo mi responsabilidad, que el Proyecto presentado con el título

Estudio de variables relevantes y análisis predictivo en la probabilidad de lesión de futbolistas en los campeonatos de Latinoamérica

en la ETS de Ingeniería - ICAI de la Universidad Pontificia Comillas en el

curso académico 2024/25 es de mi autoría, original e inédito y

no ha sido presentado con anterioridad a otros efectos.

El Proyecto no es plagio de otro, ni total ni parcialmente y la información que ha sido

tomada de otros documentos está debidamente referenciada.

Fdo.: María Ángeles Moreno Nieto Fecha: 06/ 07/ 2025

Autorizada la entrega del proyecto

EL DIRECTOR DEL PROYECTO

Fdo.: Luis Francisco Sánchez Merchante Fecha: 06/ 07/ 2025

GRADO EN INGENIERÍA EN TECNOLOGÍAS DE TELECOMUNICACIÓN

TRABAJO FIN DE GRADO

Estudio de variables relevantes y análisis predictivo en la probabilidad de lesión de futbolistas en los campeonatos de Latinoamérica

Autor: María Ángeles Moreno Nieto

Director: Luis Francisco Sánchez Merchante

Madrid

ESTUDIO DE VARIABLES RELEVANTES Y ANÁLISIS PREDICTIVO EN LA PROBABILIDAD DE LESIÓN DE FUTBOLISTAS EN LOS CAMPEONATOS DE LATINOAMÉRICA

Autor: Moreno Nieto, María Ángeles.
Director: Sánchez Merchante, Luis Francisco.
Entidad Colaboradora: ICAI – Universidad Pontificia Comillas

RESUMEN DEL PROYECTO

Este trabajo desarrolla un sistema predictivo de lesiones en el fútbol profesional latinoamericano utilizando análisis exploratorio y modelos de machine learning. Se identifican factores contextuales relevantes como el rol local/visitante, distancia recorrida o país anfitrión, y se entrenan modelos de clasificación que alcanzan buen rendimiento predictivo. Los resultados confirman las hipótesis del análisis previo y se integran en una herramienta interactiva aplicable para la toma de decisiones.

Palabras clave: Lesiones deportivas, Machine Learning, Fútbol en Latinoamérica, Predicción de riesgo, Modelos de clasificación binaria.

1. Introducción

Las lesiones en el fútbol profesional no solo afectan al rendimiento deportivo, sino que implican costes económicos y humanos. Se estima que los futbolistas profesionales sufren una media de 8,1 lesiones por cada 1.000 horas de exposición [1], con un impacto registrado de 732 millones de euros para la temporada 2023/24 en las cinco grandes ligas europeas [2], lo que ha impulsado el desarrollo de modelos predictivos para ayudar a prevenir riesgos en contextos reales de competición.

En este contexto, este trabajo analiza el riesgo de lesión en partidos oficiales de la Copa Libertadores y la Copa Sudamericana entre 2022 y 2024. A través de técnicas de machine learning, se busca identificar variables significativas y desarrollar una herramienta predictiva explicable y aplicable.

2. Definición del proyecto

Se parte de un dataset anonimizado de partidos y lesiones ocurridas en la Copa Libertadores y la Copa Sudamericana entre 2022 y 2024. Se propone identificar las variables más influyentes en la probabilidad de que ocurra una lesión durante un partido y construir modelos que permitan anticipar este riesgo. Entre las variables analizadas destacan: país anfitrión, condición de local o visitante, temperatura, altitud, distancia recorrida por el equipo visitante, fase del torneo, y tipo de competición.

Adicionalmente, se proponen hipótesis sobre el efecto de factores como el rol del equipo (local/visitante), la distancia recorrida, las condiciones geográficas o el calendario competitivo, y se exploran mediante análisis descriptivos y pruebas estadísticas.

3. Descripción del modelo

El sistema desarrollado incluye dos componentes principales:

- Un análisis exploratorio que permite identificar relaciones entre variables clave y la ocurrencia de lesiones.
- Un conjunto de modelos de clasificación (Regresión Logística, Random Forest, XGBoost y Stacking, cuya arquitectura se ve en la Figura 1) entrenados para predecir si habrá o no una lesión en un partido, basados en variables como país anfitrión, equipo, temperatura, altitud, distancia recorrida y fase del torneo.

Además, se han desarrollado dos interfaces de usuario: una que permite entrenar modelos personalizados seleccionando variables y tipo de lesión, y otra que, con los datos de un nuevo partido, ofrece la probabilidad estimada de lesión, ambas adaptables y escalables a nuevas competiciones y tipos de lesiones.

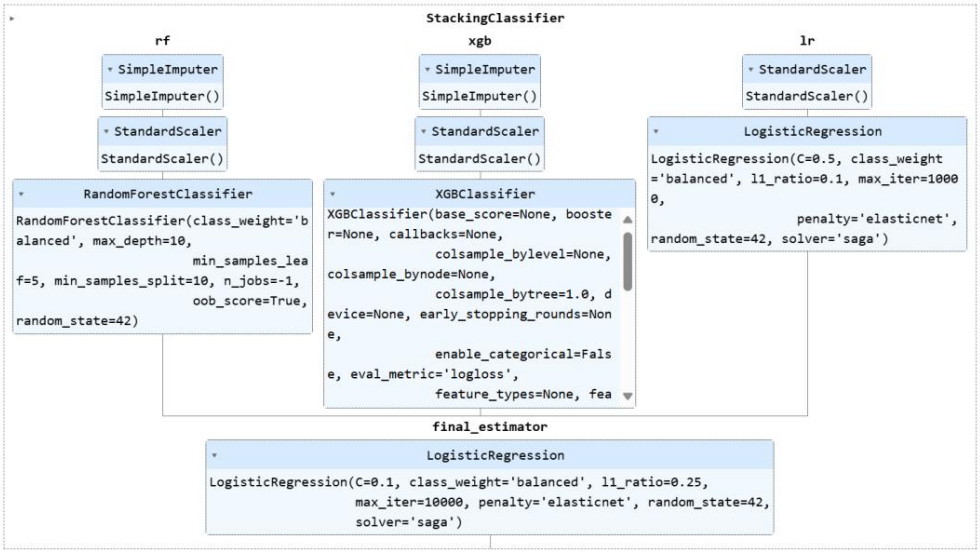

Figura 1: Arquitectura del modelo predictivo basado en stacking

4. Resultados y conclusiones

El análisis exploratorio da lugar a varias hipótesis clave: los jugadores locales sufren más lesiones, posiblemente por sobre esfuerzo o mayor agresividad del rival; el nuevo calendario con playoffs introducido en 2023 [3] ha incrementado el riesgo lesivo, especialmente para los locales que enfrentan a visitantes que han viajado más de 1.000 km. Variables como la altitud o la temperatura no muestran relación lineal directa, pero pueden tener influencia al combinar con otras variables.

A nivel predictivo, los modelos refuerzan estos hallazgos: el país anfitrión, el año del torneo, la distancia recorrida y la condición de local o visitante destacan como predictores. Este último se ve respaldado por un estudio que vincula la presencia de la afición local con un comportamiento intenso del anfitrión y agresivo del visitante [4].

Tras entrenar los diferentes tipos de modelos se evaluaron con diferentes métricas, mostrando los mejores para cada tipo en la Tabla 1. Además, el modelo que mejor capacidad de detección de lesiones reales tiene es el de stacking (que combina los tres anteriores), al tener el mayor valor de recall lo que minimiza los falsos negativos y cuyo valor se muestra en la Tabla 1.

Tipo de modelo	Accuracy	Precision	Recall	AUC	F1 - score
Reg. Logística final	0.5775	0.4193	0.6666	0.5338	0.51485
XGBoost final	0.6373	0.4151	0.3859	0.5467	0.4000
Random forest final	0.6044	0.4026	0.5439	0.5761	0.4627
Stacking con Reg. Logística	0.5862	0.4348	0.7692	0.6254	0.5556

Tabla 1: Métricas de evaluación de los 4 modelos finales

En conjunto, se han alcanzado los objetivos de realizar un análisis contextual detallado, identificar variables explicativas y desarrollar un sistema de predicción funcional. Como valor añadido, se ha implementado una herramienta interactiva que permite entrenar modelos personalizados y estimar el riesgo de lesión en partidos futuros, reforzando su aplicabilidad por parte de cuerpos técnicos y personal médico.

Este sistema representa un primer paso hacia la implementación real de modelos predictivos de lesiones en el fútbol profesional, y abre nuevas vías para incorporar variables más detalladas, adaptar el modelo a otras competiciones, y mejorar progresivamente su rendimiento con datos adicionales.

5. Referencias

[1] Barça Innovation Hub, «¿Cuánto se lesiona un jugador profesional? Un análisis de la epidemiología de las lesiones en el fútbol» 31 Mayo 2021. [En línea]. Available: https://barcainnovationhub.fcbarcelona.com/es/blog/cuanto-se-lesiona-un-jugador-profesional-un-analisis-de-la-epidemiologia-de-las-lesiones-en-el-futbol/

[2] Howden group, «Índice de lesiones fútbol 23-24 (resumen español),» 15 Octubre 2024. [En línea]. Available: https://www.howdengroup.com/es-es/reports/indice-de-lesiones-en-el-futbol-europeo-masculino-2023-24

[3] Conmebol, «Con cambios en el formato, la CONMEBOL Sudamericana gana aún más competitividad y atractivo.,» 22 Diciembre 2022. [En línea]. Available: https://www.conmebol.com/noticias/con-cambios-en-el-formato-la-conmebol-sudamericana-gana-aun-mas-competitividad-y-atractivo/

[4] F. Wunderlich, M. Weigett, R. Rein y D. Memmert, «How does spectator presence affect football? Home advantage remains in European top-class football matches played without spectators during the COVID-19 pandemic » *PLOS ONE,* 2021. 16(3): e0248590. https://doi.org/10.1371/journal.pone.0248590

STUDY OF RELEVANT VARIABLES AND PREDICTIVE ANALYSIS OF INJURY PROBABILITY IN FOOTBALL PLAYERS DURING LATIN AMERICAN CHAMPIONSHIPS

Author: Moreno Nieto, María Ángeles.
Supervisor: Sánchez Merchante, Luis Francisco.
Collaborating Entity: ICAI – Universidad Pontificia Comillas

ABSTRACT

This paper develops a predictive system for injuries in Latin American professional football using exploratory data analysis and machine learning models. It identifies key contextual factors, such as home/away status, travel distance, or host country, and trains classification models that achieve solid predictive performance. The results confirm previous hypotheses and are integrated into an interactive tool designed to support coaching and medical staff in decision-making.

Keywords: Sports Injuries, Machine Learning, Latin American Soccer, Injury Risk Prediction, Binary Classification Models

1. Introduction

Injuries in professional soccer not only impact athletic performance but also involve significant economic and human costs. It is estimated that professional footballers suffer an average of 8.1 injuries per 1,000 hours of exposure [1], with a recorded economic impact of €732 million during the 2023/24 season across the five major European leagues [2]. This has driven the development of predictive models aimed at preventing injury risks in real competitive settings.

In this context, this paper analyzes injury risk in official Copa Libertadores and Copa Sudamericana matches between 2022 - 2024. Using machine learning techniques, it aims to identify significant variables and develop an explainable and practical predictive tool.

2. Project Definition

The study is based on an anonymized dataset of matches and reported injuries from the Copa Libertadores and Copa Sudamericana between 2022 and 2024. Its goal is to identify the most influential contextual variables affecting injury risk and to construct models capable of anticipating these injuries. The variables analyzed include host country, home/away status, temperature, altitude, travel distance of the visiting team, tournament phase, and competition type.

Additionally, hypotheses about the impact of factors such as geographic conditions, team roles, or the new competitive calendar are tested through descriptive analysis and statistical testing.

3. Descripción del modelo/sistema/herramienta

The developed system consists of two main components:

- An exploratory analysis to identify relationships between key variables and injury occurrence.
- A set of classification models (Logistic Regression, Random Forest, XGBoost, and Stacking, whose architecture is shown in Figure 2) trained to predict whether an injury will occur in a given match, based on variables such as host country, temperature, altitude, team identity, travel distance, and tournament phase.

In addition, two user interfaces have been developed: one allows users to train custom models by selecting input variables and injury types, and the other estimates the probability of injury based on new match data provided by the user. Both tools are adaptable and scalable to new competitions and injury types.

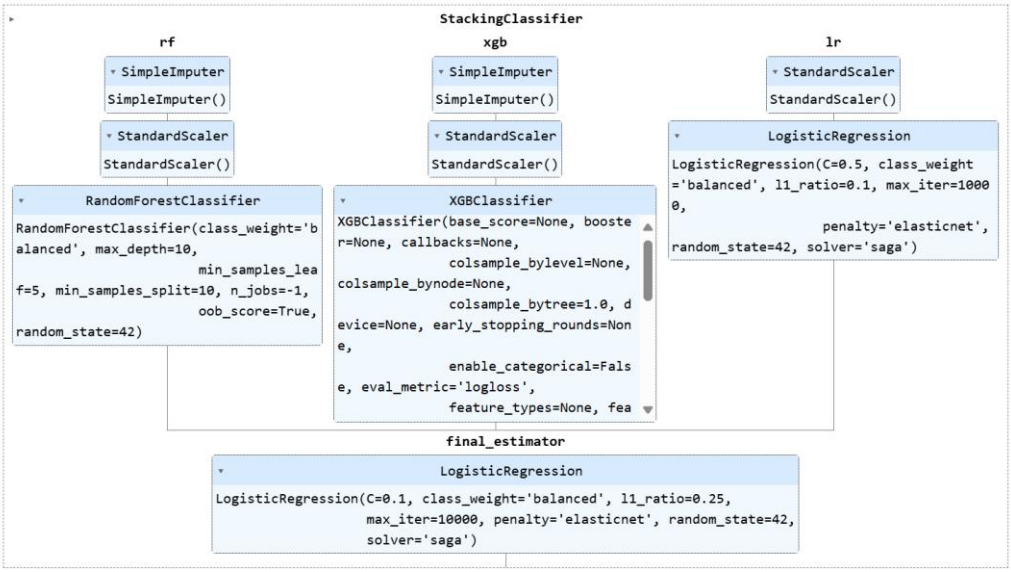

Figura 2: Predictive model architecture based on stacking

4. Results and conclusions

The exploratory analysis led to several key hypotheses: home players tend to suffer more injuries, possibly due to overexertion or increased opponent aggression; the new playoff format introduced in 2023 [3] seems to increase injury risk, particularly for home teams facing opponents who have traveled more than 1000km. Altitude and temperature do not show linear effect but are influential when combined with others.

From a predictive perspective, the models support these findings: host country, visitor travel distance, tournament year, and home/away status emerge as consistent predictors. The latter is further supported by existing research linking the presence of home fans with more intense behavior by the home team and greater aggressiveness from the visitors [4].

After training the different types of models, their performance was evaluated using various metrics. The best-performing models for each type are summarized in Table 1. Notably, the stacking model, which combines logistic regression, XGBoost, and Random Forest, achieves the best ability to detect actual injuries, as it yields the highest recall value, thereby minimizing false negatives.

Tipo de modelo	Accuracy	Precision	Recall	AUC	F1 - score
Reg. Logística final	0.5775	0.4193	0.6666	0.5338	0.51485
XGBoost final	0.6373	0.4151	0.3859	0.5467	0.4000
Random forest final	0.6044	0.4026	0.5439	0.5761	0.4627
Stacking con Reg. Logística	0.5862	0.4348	0.7692	0.6254	0.5556

Tabla 2: Performance metrics of final models

Overall, the objectives of conducting a detailed contextual analysis, identifying explanatory variables, and developing a functional prediction system have been achieved. As an added value, an interactive tool has been implemented that allows users to train customized models and estimate injury risk in future matches, enhancing its applicability for coaching and medical staff.

This system represents a first step toward the practical implementation of predictive injury models in professional football and opens new paths to incorporate more detailed variables, adapting the model to other competitions, and improving its performance with additional data.

5. References

[1] Barça Innovation Hub, «¿Cuánto se lesiona un jugador profesional? Un análisis de la epidemiología de las lesiones en el fútbol» 31 Mayo 2021. [En línea]. Available: https://barcainnovationhub.fcbarcelona.com/es/blog/cuanto-se-lesiona-un-jugador-profesional-un-analisis-de-la-epidemiologia-de-las-lesiones-en-el-futbol/

[2] Howden group, «Índice de lesiones fútbol 23-24 (resumen español),» 15 Octubre 2024. [En línea]. Available: https://www.howdengroup.com/es-es/reports/indice-de-lesiones-en-el-futbol-europeo-masculino-2023-24

[3] Conmebol, «Con cambios en el formato, la CONMEBOL Sudamericana gana aún más competitividad y atractivo.,» 22 Diciembre 2022. [En línea]. Available: https://www.conmebol.com/noticias/con-cambios-en-el-formato-la-conmebol-sudamericana-gana-aun-mas-competitividad-y-atractivo/

[4] F. Wunderlich, M. Weigett, R. Rein y D. Memmert, «How does spectator presence affect football? Home advantage remains in European top-class football matches played without spectators during the COVID-19 pandemic » *PLOS ONE,* 2021. 16(3): e0248590. https://doi.org/10.1371/journal.pone.0248590

Índice de la memoria

Índice de figuras

Índice de tablas

Capítulo 1. INTRODUCCIÓN

El fútbol profesional es una de las industrias deportivas más influyentes a nivel global, no solo por su impacto cultural y mediático, sino también por el volumen económico que moviliza cada temporada. Sin embargo, más allá del espectáculo y la competición, uno de los retos más persistentes y costosos a los que se enfrentan los clubes es la gestión de las lesiones deportivas. Estas no solo condicionan el rendimiento de los jugadores, sino que afectan directamente a la competitividad de los equipos y a su sostenibilidad económica.

En promedio, los futbolistas profesionales sufren alrededor de 8,1 lesiones por cada 1.000 horas de exposición, siendo más frecuentes durante los partidos (36 lesiones por cada 1.000 horas) que en los entrenamientos (3,7 por cada 1.000 horas) [1]. Este problema no es una novedad, casos como el de Victor Valdés, portero clave del FC Barcelona, que tras una rotura de ligamento cruzado en 2014 se perdió el Mundial de Brasil, o el de Ronaldo Nazário, uno de los mejores delanteros de la historia, que sufrió una doble rotura del tendón rotuliano limitando permanentemente su potencia física, ilustran como una lesión puede marcar un antes y un después en la carrera de un jugador [2].

Además del impacto deportivo, las lesiones generan un elevado coste económico. Fichajes millonarios que pasan largos periodos sin competir reducen drásticamente el retorno de inversión esperado por los clubes. Casos como los de Eden Hazard en el Real Madrid u Ousmane Dembélé en el FC Barcelona muestran cómo una elevada inversión económica puede traducirse en bajo rendimiento deportivo debido a reiteradas lesiones musculares [3].

Según el informe de Howden Group, durante la temporada 2023/24, las cinco principales ligas europeas registraron un total de 4.123 lesiones, suponiendo un aumento del 4% en la frecuencia de lesiones y un incremento del 5% en los costos asociados en comparación con la temporada anterior. Estos gastos alcanzaron los 732,02 millones de euros, siendo la Premier League inglesa la más afectada, con un total de 318,8 millones de euros en gastos relacionados con lesiones, representando el 44% del gasto total [4].

Si bien estos datos corresponden a Europa, en Latinoamérica el problema puede ser incluso más crítico debido a la menor disponibilidad de recursos. Además, las competiciones como la Copa Libertadores y la Copa Sudamericana se disputan bajo condiciones geográficas extremas (altitud, calor, humedad) y calendarios muy exigentes, lo que incrementa el riesgo de lesión [5].

Ante este escenario, la ciencia aplicada al deporte ha avanzado notablemente en los últimos años. Se han incorporado herramientas como el análisis biomecánico y los modelos predictivos basados en inteligencia artificial (IA) para reducir el riesgo de lesiones. Instituciones como el Instituto de Biomecánica de Valencia (IBV) [6] y clubes de élite como el Manchester City han implementado soluciones que analizan patrones de movimiento y datos de entrenamiento para personalizar la carga física y prevenir lesiones [7]. Tecnologías como las de Catapult Sports combinan sensores GPS y algoritmos para monitorizar en tiempo real la fatiga y recuperación del jugador. [8].

No obstante, los factores contextuales como la altitud, el clima, el rol del equipo (local o visitante), o el diseño del calendario siguen poco explorados, pese a que pueden desempeñar un papel clave. En el caso de Latinoamérica, donde las competiciones se disputan en entornos extremos y bajo calendarios saturados, el análisis de estas condiciones cobra especial relevancia ya que existen estudios que demuestran que factores como la altitud y el calor pueden afectar significativamente el rendimiento de los jugadores [9].

Este Trabajo surge de la necesidad de analizar de forma integral estos factores contextuales por el equipo del Dr. Francisco Forriol, que incluye lesiones registradas entre 2022 y 2024 en las principales competiciones del continente. Esta información se ha enriquecido con variables geográficas, climáticas y de calendario, con el objetivo de construir un modelo predictivo capaz de anticipar el riesgo de lesión según el contexto del partido.

El enfoque adoptado combina análisis exploratorio, tratamiento de datos faltantes y construcción de modelos de clasificación mediante algoritmos como regresión logística,

Random Forest, XGBoost y técnicas de ensemble. Asimismo, se ha desarrollado una interfaz interactiva que permite adaptar el sistema a nuevas ligas o tipos de lesión, favoreciendo su aplicación en entornos reales.

En definitiva, este trabajo tiene como objetivo principal contribuir al campo de la ingeniería aplicada al deporte mediante el desarrollo de una herramienta práctica para cuerpos técnicos que facilite la toma de decisiones informadas y adaptadas al entorno latinoamericano., permita anticipar el riesgo de lesiones en el fútbol profesional.

Los capítulos siguientes desarrollan en detalle los aspectos metodológicos, tecnológicos y analíticos de esta investigación, desde la limpieza y el análisis exploratorio de los datos, hasta la construcción de los modelos y su evaluación en un entorno interactivo

Capítulo 2. DESCRIPCIÓN DE LAS TECNOLOGÍAS

El proyecto ha sido íntegramente desarrollado en **Python**, un lenguaje de programación con sintaxis sencilla y librerías especializadas que lo convierten en una buena herramienta para el análisis de datos y la implementación de modelos de machine learning. Para su implementación se ha utilizado el entorno **Jupyter Notebook**.

Para el tratamiento, limpieza y transformación del conjunto de datos original, se ha recurrido al siguiente conjunto de librerías:

- **pandas**: herramienta central para la manipulación de estructuras tipo DataFrame. Se ha utilizado para la carga, depuración, filtrado condicional, agrupación, generación de nuevas variables y operaciones estadísticas básicas [10].
- **numpy**: soporte de estructuras numéricas eficientes, operaciones vectorizadas y cálculos auxiliares requeridos por otras librerías [11].
- **datetime y calendar**: empleados para la creación de variables temporales derivadas (mes, trimestre, días desde el 1 de enero) [12].
- **Re** (expresiones regulares): utilizada en la normalización de cadenas de texto, especialmente en la estandarización de nombres de equipos y países [13].
- **openpyxl** [14] y **xlsxwriter** [15]: para la lectura y escritura de archivos Excel durante la fase de limpieza, así como la creación del dataset final.

Durante la fase de análisis exploratorio y visualización de relaciones entre variables, se han empleado las siguientes:

- **matplotlib** [16] y **seaborn** [17]: utilizadas para la generación de gráficos estáticos como histogramas, diagramas de barras, box plots, mapas de calor y gráficos de dispersión.

Finalmente, en la fase de modelado y evaluación de la variable lesión y la posterior predicción de las variables separadas por diagnóstico y localización, se ha hecho uso de las siguientes:

- **scikit-learn:** herramienta para el entrenamiento y validación de modelos supervisados como regresión logística, árboles de decisión, Random Forest y modelos de stacking. También ha sido utilizada para el escalado de variables (StandardScaler), la selección de hiper parámetros (GridSearchCV), y el cálculo de métricas como precisión, recall, F1-score, curva ROC y AUC [18].
- **xgboost:** librería especializada en boosting, muy eficiente para problemas de clasificación estructurada con datos tabulares [19].
- **imbalanced-learn:** para el tratamiento de desbalanceo de clases [20].
- **ipywidgets**: para la herramienta final dirigida al usuario ya que permite crear una interfaz gráfica basada en selección de parámetros, filtrado de columnas y visualización de resultados [21].

Capítulo 3. ESTADO DE LA CUESTIÓN

El creciente impacto económico y deportivo de las lesiones en el fútbol profesional ha impulsado en los últimos años el desarrollo de soluciones tecnológicas para su prevención. Este capítulo revisa los principales trabajos realizados en esta línea para entender el estado actual de la cuestión, identificar vacíos y enmarcar la contribución del proyecto.

3.1 TECNOLOGÍAS APLICADAS AL ANÁLISIS FÍSICO Y BIOMECÁNICO

3.1.1 DISPOSITIVOS GPS Y ANÁLISIS BIOMECÁNICO

En los últimos años, el uso de dispositivos de monitoreo GPS (o wearables) en el fútbol profesional se ha convertido en una herramienta clave para analizar el rendimiento de los jugadores y prevenir lesiones. Estos **dispositivos permiten registrar variables como la distancia recorrida, la velocidad, los esprints o los cambios de dirección**, proporcionando información objetiva para **ajustar la carga de trabajo y reducir el riesgo de sobre entrenamiento**.

Catapult Sports, por ejemplo, combina sensores GPS con acelerómetros triaxiales para medir en tiempo real el esfuerzo físico de cada jugador. Una de sus métricas destacadas es el PlayerLoad™, que resume la carga total del jugador a partir de los movimientos en los tres ejes espaciales, facilitando la personalización del entrenamiento y la prevención de lesiones musculares [22].

De forma similar**, Oliver Sports** ha desarrollado un dispositivo que, además de registrar métricas físicas como la velocidad máxima o las aceleraciones, incorpora inteligencia artificial para estimar un índice de riesgo de lesión basado en los datos recogidos. Según la empresa, esta tecnología ha logrado reducir hasta en un 45 % la incidencia de lesiones musculares en los equipos que la emplean [23].

Además, equipos como el **Sevilla FC** o el **Villarreal CF** ya emplean esta tecnología para monitorizar individualmente a sus jugadores, ajustar la planificación física y prevenir recaídas. Gracias a su integración con aplicaciones móviles y plataformas de análisis, los cuerpos técnicos pueden personalizar los entrenamientos y tomar decisiones basadas en evidencia [7].

3.1.2 ANÁLISIS BIOMECÁNICO Y TÉCNICAS DE EVALUACIÓN DEL GESTO

El análisis biomecánico **estudia detalladamente los movimientos del cuerpo humano** para identificar patrones que puedan predisponer a lesiones. Investigaciones han demostrado que una **correcta técnica de carrera reduce** significativamente el riesgo de **lesiones musculares**, y el análisis en 3D del movimiento de cadera, rodilla, tobillo y pie permite detectar desequilibrios y corregirlos antes de que se conviertan en lesiones [24].

Mediante el uso de tecnologías avanzadas, como sensores inerciales y escáneres 3D, es posible evaluar la técnica de carrera, el gesto deportivo y la distribución de cargas en las diferentes articulaciones.

Podoactiva, empresa especializada en podología y biomecánica, ha implementado soluciones como el sistema de sensores inerciales (IMUs), que permite analizar la cinemática articular de la cadera, rodilla y tobillo en los tres ejes del espacio [25]. Estos sensores, proporcionan datos sobre movimientos articulares como flexión, extensión, rotación y abducción, facilitando la detección de desequilibrios y asimetrías que podrían derivar en lesiones. Además, su escáner 3D Scan Sport, patentado a nivel mundial, permite diseñar plantillas personalizadas mediante impresión 3D, adaptadas a las necesidades de cada deportista [26].

Por otro lado, el **Instituto de Biomecánica de Valencia (IBV)** ha desarrollado soluciones que combinan análisis biomecánico e inteligencia artificial para mejorar la salud y el rendimiento. Sus algoritmos de aprendizaje automático permiten evaluar el movimiento humano sin necesidad de marcadores corporales, lo que facilita un análisis más natural y preciso en condiciones reales de entrenamiento o competición [6].

3.2 MODELOS PREDICTIVOS Y HERRAMIENTAS BASADAS EN IA

3.2.1 APRENDIZAJE AUTOMÁTICO APLICADO AL RIESGO DE LESIÓN

El uso de inteligencia artificial (IA) y aprendizaje automático (machine learning) ha transformado el enfoque hacia la predicción y prevención de lesiones. A partir del análisis de grandes volúmenes de datos, como el historial médico, la carga de entrenamiento o las métricas de rendimiento, es posible identificar patrones que preceden a una lesión y anticipar su ocurrencia.

Proyectos como **SoccerGuard** han desarrollado modelos específicos para fútbol femenino profesional, integrando datos de múltiples fuentes como sensores GPS, informes de bienestar, estadísticas de rendimiento y registros médicos. Los resultados muestran que, con las configuraciones adecuadas y la combinación óptima de características, es posible predecir eventos de lesiones con una precisión considerable. Además, el sistema cuenta con una interfaz gráfica que permite la visualización interactiva de los datos y resultados [27].

En España, clubes como el **Marbella FC** han implementado soluciones basadas en IA en colaboración con la empresa Olocip. Esta plataforma centraliza datos médicos y de rendimiento en un Data Lake, permitiendo su recopilación, transformación y visualización automatizada. El sistema facilita el seguimiento de la salud de los jugadores, la detección de posibles anomalías y la evaluación de riesgos, incorporando además estimaciones de impacto deportivo y económico para apoyar la toma de decisiones técnicas y médicas [28].

3.2.2 INTEGRACIÓN DE PLATAFORMAS EN CLUBES PROFESIONALES

La integración de herramientas predictivas basadas en IA en clubes de élite ha demostrado ser eficaz para reducir el número de lesiones y mejorar la preparación física. Un caso representativo es el del **Manchester City**, que ha adoptado sistemas inteligentes para personalizar entrenamientos y aplicar estrategias preventivas basadas en datos en tiempo real.

Estos algoritmos analizan variables como el historial de lesiones, el nivel de fatiga o la calidad del sueño, permitiendo ajustar las cargas de trabajo y prevenir sobreesfuerzos. Según Rodrigo Vivares, director regional de **IPG Mediabrands**, este enfoque ha contribuido a una reducción del 20–30% en la tasa de lesiones respecto a métodos tradicionales [29].

Además, estas plataformas permiten a los entrenadores tomar decisiones tácticas en tiempo real durante los partidos, como la sustitución de un jugador en riesgo de lesión, basándose en las recomendaciones generadas por la IA. Este enfoque ha sido clave para optimizar el rendimiento y proteger la salud de jugadores de alto nivel como **Kevin De Bruyne** [29].

3.3 SOLUCIONES COMERCIALES Y HERRAMIENTAS EMERGENTES

3.3.1 TERMOGRAFÍA INFRARROJA

La termografía infrarroja es una técnica no invasiva que mide la temperatura de la superficie corporal y genera mapas térmicos que detecta asimetrías. En deporte profesional, se utiliza para identificar desequilibrios térmicos entre extremidades o zonas corporales, que pueden estar asociados a procesos inflamatorios o disfunciones musculares.

Según la empresa **ThermoHuman**, esta tecnología puede detectar alteraciones incluso antes de que el deportista experimente dolor u otros síntomas, permitiendo intervenir de forma preventiva. En el caso del fútbol, se han observado patrones térmicos repetitivos previos a lesiones, lo que permite ajustar cargas de trabajo, iniciar tratamientos tempranos o incluso retirar a un jugador de una convocatoria como medida de precaución Su facilidad de aplicación, menos de 30 segundos y sin contacto físico, la convierte en una herramienta eficiente y adaptable al entorno competitivo [30].

3.3.2 SISTEMAS DE ANÁLISIS DE VIDEO

El análisis de vídeo ha evolucionado más allá de la observación visual, integrando grandes volúmenes de datos para el estudio detallado del rendimiento. Una de las herramientas más completas es **Mediacoach**, desarrollada por La Liga en colaboración con el proyecto **Beyond Stats**, que proporciona más de 3.500 variables por partido.

Mediacoach permite analizar zonas de influencia de cada jugador, la presión ejercida, la progresión de jugadas, las coberturas defensivas o las trayectorias individuales durante todo el partido, basándose en sistemas de tracking que recogen información de los movimientos de los jugadores en el campo con alta precisión. Esta información se convierte en inteligencia táctica y física que facilita la toma de decisiones tanto a corto plazo (durante el partido) como en la planificación física y la prevención de la fatiga y lesiones [31].

3.4 VACÍOS DETECTADOS Y OPORTUNIDAD DE INVESTIGACIÓN

A pesar de los avances tecnológicos y científicos en la prevención de lesiones en el fútbol profesional, todavía persisten vacíos significativos:

- **Enfoque limitado a variables internas**: La mayoría de las soluciones actuales se centran en variables internas del jugador, como la carga de entrenamiento y el historial médico, dejando de lado factores contextuales como la altitud del estadio, el clima local, si el jugador es local o visitante, o la acumulación de partidos.
- **Falta de adaptabilidad a contextos específicos:** Muchas de las plataformas comerciales están pensadas para clubes con alto presupuesto y estructuras profesionales consolidadas, como los de las grandes ligas europeas. Esta falta de adaptación a las condiciones específicas de las competiciones latinoamericanas limita su aplicabilidad.
- **Ausencia de modelos explicativos interpretables:** Gran parte de las soluciones basadas en inteligencia artificial funcionan como cajas negras, dificultando la comprensión de por qué un jugador tiene mayor riesgo de lesión.

La identificación de estos vacíos justifica la necesidad de un enfoque alternativo, como el que propone este proyecto, que combine rigor técnico, interpretabilidad y aplicabilidad real.

Capítulo 4. DEFINICIÓN DEL TRABAJO

4.1 JUSTIFICACIÓN

El análisis de soluciones existentes realizado en el capítulo anterior ha dejado clara una limitación estructural en la mayoría de las propuestas actuales orientadas a la prevención de lesiones en el fútbol profesional. Estas soluciones, basadas en tecnologías como sensores GPS, análisis biomecánico y modelos de machine learning, se centran en variables internas del jugador y están diseñadas para contextos con alta capacidad tecnológica y presupuestaria, como las grandes ligas europeas [8].

Sin embargo, en competiciones latinoamericanas como la Copa Libertadores o la Copa Sudamericana, los equipos enfrentan retos únicos: altitudes extremas, climas variables, desplazamientos largos y frecuentes, infraestructura médica desigual y, en muchos casos, recursos económicos limitados. Pese a que estas condiciones incrementan el riesgo lesivo, no existen herramientas del mercado adaptadas a este contexto con una lógica de bajo coste, aplicabilidad real y enfoque contextual.

Este proyecto surge con el propósito de demostrar que es posible construir una herramienta útil para la prevención de lesiones basada en datos fácilmente disponibles, con modelos interpretables y centrada en variables que reflejan las condiciones reales de los clubes latinoamericanos.

Desde una perspectiva práctica, la solución propuesta no requiere dispositivos externos ni suscripciones a plataformas. Está diseñada para ser utilizada por entrenadores o preparadores físicos con formación técnica básica y permite ajustar la carga de entrenamiento o rotaciones a partir de datos contextuales.

4.2 OBJETIVOS

El objetivo principal es analizar e identificar las variables que influyen significativamente en la aparición de lesiones en el fútbol profesional latinoamericano, con especial atención a las competiciones de la Copa Libertadores y la Copa Sudamericana. Para ello se han descrito los siguientes objetivos generales:

- **Analizar descriptivamente el conjunto de datos**, para identificar patrones relacionados con la aparición de lesiones en competiciones latinoamericanas.
- **Evaluar el impacto de variables contextuales clave**, como la altitud del estadio, la condición de local o visitante y la acumulación de partidos, con el objetivo de determinar en qué medida influyen sobre el riesgo lesivo.
- **Desarrollar modelos predictivos explicables**, utilizando técnicas estadísticas y algoritmos de machine learning para estimar la probabilidad de que ocurra una lesión en función de distintas combinaciones de variables.
- **Implementar herramientas** prácticas basadas en dichos modelos, incluyendo una interfaz interactiva, que permita su utilización por parte de usuarios no expertos y facilite la toma de decisiones en contextos reales

Además, también se han identificado los siguientes objetivos específicos:

- **Identificar variables individuales con mayor peso explicativo**, incluso en los casos en los que el modelo general no resulte concluyente, con el fin de destacar factores de riesgo concretos.

- **Realizar análisis por subgrupos**, dividiendo los datos según el tipo de lesión (por ejemplo, muscular, traumática) o la localización anatómica afectada, para descubrir patrones más precisos.

4.3 METODOLOGÍA

La metodología adoptada sigue una estructura por fases sucesivas, orientadas a cumplir los objetivos anteriores y combinando las técnicas de análisis estadístico, aprendizaje automático y visualización interactiva.

4.3.1 LIMPIEZA Y PREPARACIÓN DEL DATASET

Se parte del conjunto de datos suministrado por el Dr. Francisco Forriol, que incluye registros de lesiones en partidos de la Copa Libertadores y Sudamericana. Se realiza una limpieza y normalización exhaustiva: eliminación de duplicados, homogeneización de nombres de equipos y tratamiento de valores nulos.

Con el objetivo de capturar mejor el contexto en el que se producen las lesiones, se amplía el conjunto de datos con las siguientes variables:

- **Días transcurridos desde el 1 de enero**, a partir de la fecha del partido, para aproximar el momento de la temporada y relacionarlo con el desgaste acumulado.
- **Fase agrupada del torneo**, obtenida reagrupando las distintas fases del campeonato (como fase de grupos, octavos, cuartos, etc.) en categorías más generales que reflejan el nivel de exigencia competitiva.
- **Diagnósticos agrupados**, a partir de la variable original para simplificar el análisis por categorías clínicas.
- **Localizaciones agrupadas**, simplificando las múltiples localizaciones anatómicas.

4.3.2 ANÁLISIS EXPLORATORIO DE LOS DATOS

Una vez consolidado el conjunto de datos final, se lleva a cabo un análisis exploratorio en dos fases. En la primera, se realiza un estudio descriptivo de todas las variables que conforman el dataset, con el objetivo de comprender mejor el contexto en el que se producen las lesiones. En la segunda, se analiza la relación entre estas variables y la aparición de

lesiones, con el fin de detectar patrones iniciales, establecer comparaciones entre grupos y formular hipótesis preliminares.

4.3.3 ENTRENAMIENTO DE MODELOS PREDICTIVOS

Primero se define la variable binaria lesión que representa la aparición o no de lesión en un partido. Esto permite entrenar los siguientes modelos de clasificación: Regresión logística (con penalización ElasticNet), Random Forest, XGBoost, modelos de combinación como votación ponderada y Stacking.

La evaluación incluye validación cruzada, ajuste de hiper parámetros (GridSearchCV) y análisis de métricas como accuracy, precision, recall, F1-score y AUC. Además se ajuste el umbral, para maximizar el recall, priorizando la detección de lesiones reales.

4.3.4 INTERPRETACIÓN DE RESULTADOS Y SELECCIÓN DE VARIABLES CLAVE

Una vez entrenados los modelos y evaluado su rendimiento, se realiza un análisis de las variables más relevantes, analizando la consistencia entre los clasificadores y permitiendo confirmar hipótesis generadas en el análisis exploratorio.

Este paso es clave tanto para validar los modelos como para extraer conocimiento práctico aplicable en el ámbito deportivo. Además, se explora la posibilidad de especializar los modelos para predecir no solo si ocurrirá una lesión, sino también predecir por tipo de lesión, siempre que la cantidad de datos disponibles en cada categoría lo permita.

Finalmente se desarrolla una interfaz interactiva en Jupyter Notebook con ipywidgets, para permitir a los usuarios: Seleccionar el tipo de predicción (general, por diagnóstico o por localización), incluir o excluir variables contextuales según el caso, elegir entre los modelos predictivos entrenados y ejecutar una predicción de un partido concreto para ver su probabilidad de lesión. Esto permite en tiempo real entrenar modelos ajustando a las características concretas de condiciones específicas que puedan surgir y obtener resultados útiles y aplicables.

4.4 PLANIFICACIÓN DE LAS TAREAS

La planificación del proyecto se ha distribuido entre diciembre y junio, según las tareas mostradas en la Figura 3. Cada fase ha seguido un orden lógico, desde la limpieza inicial hasta la elaboración final de la memoria:

- **Familiarización y limpieza de datos:** primera toma de contacto con el dataset, revisión del formato, eliminación de duplicados y corrección de errores.
- **Búsqueda de variables adicionales relevantes:** ampliación del dataset con variables externas como altitud, temperatura, distancia o fase del torneo.
- **Análisis descriptivo de las variables:** análisis exploratorio general y relacional para formular hipótesis e identificar tendencias iniciales.
- **Desarrollo del modelo predictivo general:** creación y evaluación de modelos base para la variable lesión.
- **Desarrollo del modelo por subconjuntos:** filtrado y modelado específico por diagnóstico y localización, con criterios de frecuencia mínima y selección de variables.
- **Interpretación de resultados y selección de variables clave:** análisis de importancia de variables y validación de hipótesis.
- **Implementación de la interfaz interactiva:** creación de una herramienta en Jupyter Notebook para predecir lesiones según el contexto del partido.
- **Elaboración del TFG:** redacción y edición del documento final, incluyendo referencias, figuras, apéndices y anexos.

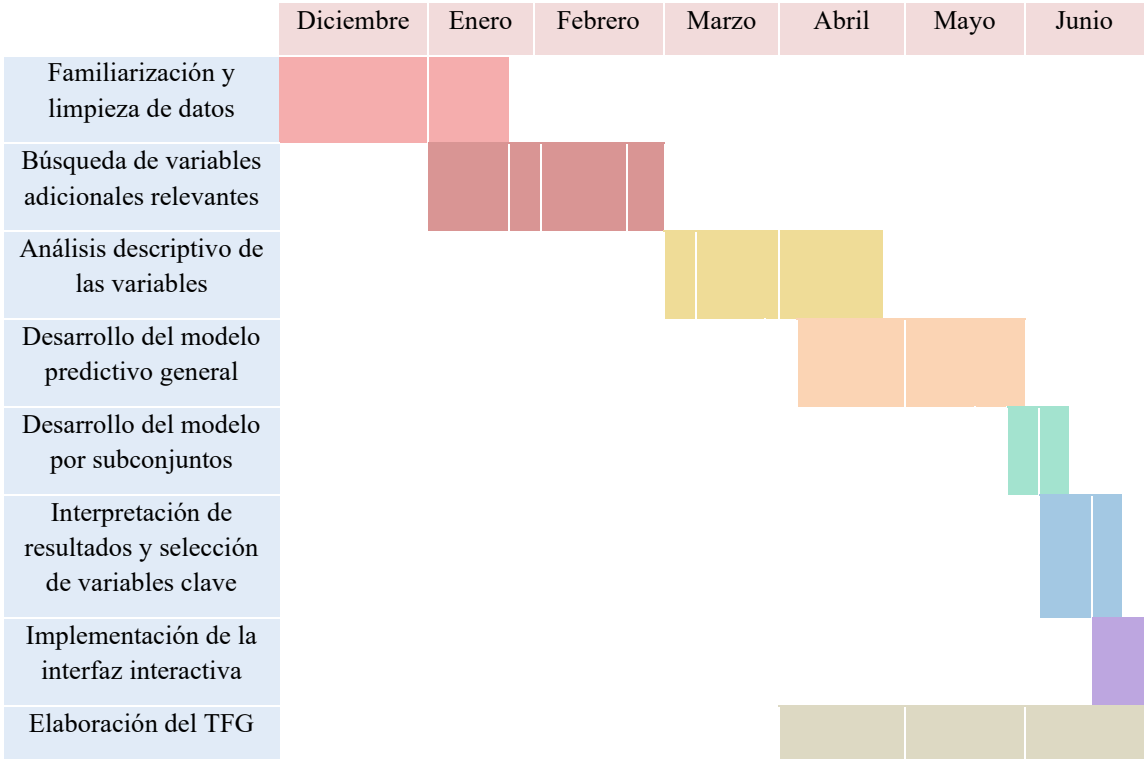

Figura 3: Gantt de planificación temporal

4.5 ANÁLISIS ECONÓMICO

La estimación económica del proyecto contempla el tiempo dedicado, los recursos utilizados y las herramientas aplicadas. Dado el carácter académico del trabajo, se han utilizado recursos propios y software libre, minimizando los costes. El desglose de los costes incurridos se muestra en la Tabla 3 y se ha calculado un coste total aproximado de 2700€.

Concepto	**Detalle técnico**	**Coste estimado**
Tiempo de trabajo	250 horas de dedicación, estimando 10 €/hora como sueldo de un ingeniero junior [32]	2500€
Amortización de equipo personal	Ordenador de 1.200 €, amortizado en 5 años, uso dedicado del 20% este año	60€
Software y librerías	Herramientas de uso libre y gratuito (Python, scikit-learn, pandas, matplotlib, etc.)	0€
Base de datos	Cedida con fines académicos por el Dr. Francisco Forriol	0€
Asistencia técnica de ChatGPT Plus	Suscripción mensual a ChatGPT Plus durante 7 meses (20 €/mes)	140€
Coste total estimado del proyecto		**2700€**

Tabla 3: Costes del proyecto

Capítulo 5. ANÁLISIS EXPLORATORIO DE LOS DATOS

5.1 CONTEXTUALIZACIÓN DE LAS COMPETICIONES Y ORIGEN DEL DATASET

5.1.1 EL FÚTBOL LATINOAMERICANO Y SU CONTEXTO COMPETITIVO

Durante los años 2022, 2023 y 2024, el fútbol latinoamericano ha evolucionado en un escenario marcado por diversos retos. A pesar de haber superado los momentos críticos de la pandemia de COVID-19, sus efectos residuales todavía se dejan sentir en aspectos como la preparación física de los jugadores y la planificación de calendarios. Además, la intensa carga de partidos, los desplazamientos frecuentes entre países, las condiciones ambientales extremas y la salida constante de talento joven hacia ligas europeas han configurado un panorama de alta exigencia para los clubes y los futbolistas de la región.

Entre los factores clave que influyen en el contexto de las competiciones latinoamericanas durante este periodo destacan:

- **La congestión del calendario**: con fases clasificatorias, torneos nacionales e internacionales superpuestos, los equipos afrontan periodos con muy poco descanso entre partidos, lo que incrementa el riesgo de fatiga y sobrecarga muscular. En 2024, FIFPRO y la Asociación Mundial de Ligas denunciaron públicamente que el calendario de competiciones estaba roto debido a la acumulación excesiva de partidos, lo que compromete la salud física y mental de los jugadores [33].

- **Condiciones geográficas y climáticas extremas**: la altitud de ciudades como La Paz (más de 3.600 metros) o Quito (2.800 metros), así como las temperaturas extremas del verano en países como Brasil o Argentina, crean un entorno fisiológicamente exigente para los jugadores. Tal como explica el Sports Science Institute de

Gatorade, estas condiciones imponen una demanda cardiovascular adicional y pueden incrementar el riesgo de lesiones musculares y calambres [34].

- **La variabilidad de infraestructuras y recursos médicos**: no todos los equipos participantes disponen de las mismas condiciones de preparación, recuperación o tratamiento, lo que puede influir tanto en la aparición y manejo de lesiones.

Este contexto justifica el interés por analizar las lesiones en el fútbol latinoamericano desde una perspectiva contextualizada y adaptada a las particularidades de la región.

5.1.2 COMPETICIONES ANALIZADAS: COPA LIBERTADORES Y COPA SUDAMERICANA

- **La Copa Libertadores,** considerada el equivalente latinoamericano de la UEFA Champions League, es el torneo más prestigioso a nivel de clubes del continente. Participan 47 equipos de los diez países latinoamericanos (Argentina, Bolivia, Brasil, Chile, Colombia, Ecuador, Paraguay, Perú, Uruguay y Venezuela). El torneo se estructura en tres fases preliminares eliminatorias, una fase de grupos compuesta por 32 equipos, y rondas finales de eliminación directa (octavos, cuartos, semifinal y final). La final se disputa a partido único en sede neutral [35].

- **La Copa Sudamericana,** por su parte, es el segundo torneo en importancia a nivel de clubes en Sudamérica, similar en jerarquía a la UEFA Europa League. Participan 56 equipos de los diez países de la región. El torneo se inicia con una fase preliminar nacional (con cruces entre equipos del mismo país) en la que participan 32 equipos, una fase de grupos con los 16 clasificados de la fase anterior a los que se le suman 6 equipos de Argentina y Brasil clasificados automáticamente y los 4 equipos eliminados de la última fase de grupos de la Copa Libertadores, y una fase playoffs. Finalmente se clasifican 8 equipos para jugar los cuartos de final, semifinales y la final [36].

Ambas competiciones se disputan entre los meses de enero y noviembre, y presentan altos niveles de competitividad, viajes transcontinentales frecuentes y diversidad de escenarios (tanto en tipo de estadio como en condición climática).

5.1.3 ORIGEN Y ESTRUCTURA GENERAL DEL DATASET

El dataset utilizado en este estudio ha sido proporcionado por el equipo médico del Dr. Francisco Forriol. Incluye registros de partidos disputados en las ediciones 2022, 2023 y 2024 de la Copa Libertadores y la Copa Sudamericana. La información contenida abarca aspectos competitivos (fecha, fase, equipos), geográficos (ciudad, país, altitud), ambientales (temperatura), y datos médicos relativos a lesiones (número total, tipo de diagnóstico, localización anatómica, minuto del partido en el que ocurrió la lesión etc.).

Esta base de datos ha sido ampliada, limpiada y estructurada a lo largo de varias fases, con el objetivo de convertirla en una herramienta robusta para el análisis estadístico y predictivo del riesgo de lesión. Los detalles del proceso de limpieza se desarrollan a continuación.

5.2 LIMPIEZA DEL DATASET

5.2.1 LIMPIEZA FASE 1: HOMOGENEIZACIÓN Y ESTANDARIZACIÓN ESTRUCTURAL

En la primera etapa del proceso de limpieza se abordaron inconsistencias de formato, nombres y estructura general del dataset. Las principales acciones realizadas fueron:

- Eliminación de columnas vacías o irrelevantes, como la columna NÚMERO, completamente nula.
- Conversión de texto a mayúsculas y eliminación de acentos para asegurar uniformidad.
- Normalización de nombres de equipos mediante una tabla de correspondencias personalizada.

- Introducción a mano de la fase de algunos partidos que faltaba, buscando los partidos necesarios según su fecha en internet y encontrado su fase.
- Extracción de variables desde campos combinados: desde la columna PARTIDO, que incluía ambos equipos y países, se crearon las variables TEAM1, TEAM2, TEAM1_NAME, TEAM2_NAME, TEAM1_COUNTRY y TEAM2_COUNTRY.
- Corrección y unificación de fases del torneo: las diferentes formas en que se representaban fases como "3 IDA", "FG FECHA 2" o "CUARTOS – VUELTA" se agruparon en un campo único estandarizado (FASE_PARTIDO).

5.2.2 LIMPIEZA FASE 2: CREACIÓN DE VARIABLES CONTEXTUALES

En una segunda etapa, se generaron variables derivadas con un enfoque más analítico y orientado al estudio del riesgo de lesiones. Entre los principales cambios destacan:

- Cálculo de la variable DIAS_TRANSCURRIDOS, que representa el número de días entre el 1 de enero y la fecha del partido, como métrica de la "altura de temporada".
- Clasificación binaria IDA_VUELTA, que permite identificar si el partido corresponde a la ida, la vuelta o una final a partido único.
- Variables derivadas mediante agrupación de valores categóricos: como los campos FASE_AGRUPADA, DIAGNOSTICO_AGRUPADO y LOCALIZACION_AGRUPADA, que reducen el nivel de especificidad de los datos.

Además, para facilitar la claridad del análisis, todas las variables nuevas creadas manualmente o por transformación se han nombrado con letras en mayúsculas, de forma que se distingan de las variables originales del dataset. Estos cambios se ilustran en la Figura 4:Tabla resumen con las columnas añadidas o derivadas , donde se muestra el conjunto completo resultante tras el proceso de limpieza.

.

Columna	Descripción	Sustituye
PAIS_PARTIDO	País en el que se ha jugado el partido. Se obtiene a partir de la columna original Pais , normalizado sin tildes y en mayúsculas.	sustituye a Pais
CIUDAD_PARTIDO	Ciudad donde se ha disputado el partido. Equivale a la columna Ciudad , pero renombrada para mayor claridad.	Sustituye a Ciudad
FASE_PARTIDO	Nombre original de la fase del torneo en la que se juega el partido (ej. "Primera fase", "Fase de grupos").	Sustituye a Fase
IDA_VUELTA	Indicador que especifica si el partido es de ida o vuelta, cuando esa información está disponible.	Nueva columna
FASE_AGRUPADA	Fase agrupada en categorías generales como "FASES DE CLASIFICACION" o "FASE DE GRUPOS", derivada de FASE_PARTIDO .	Derivada de FASE_PARTIDO
TEAM1_NAME_NORM	Nombre del equipo local (Team1) normalizado (mayúsculas, sin errores tipográficos ni variantes regionales).	Derivada de Team1_Name
TEAM2_NAME_NORM	Nombre del equipo visitante (Team2) normalizado.	Derivada de Team2_Name
TEAM1_NAME_FINAL	Nombre definitivo y desambiguado del equipo local. Incluye el país entre paréntesis en caso de homónimos (ej. "RIVER PLATE (ARG)").	Derivada de TEAM1_NAME_NORM
TEAM2_NAME_FINAL	Nombre definitivo y desambiguado del equipo visitante. Incluye el país si hay equipos con nombres repetidos.	Derivada de TEAM2_NAME_NORM

Figura 4:Tabla resumen con las columnas añadidas o derivadas

5.2.3 RESULTADO FINAL DEL DATASET

Como resultado del proceso de limpieza y transformación, se ha obtenido un dataset final, cuya información se visualiza en la Figura 5: Columnas del dataset limpiado, compuesto por 908 partidos con información verificada y estructurada y más de 30 variables entre numéricas, categóricas y binarias.

Este dataset constituye una herramienta sólida y fiable para investigar las relaciones entre el contexto del partido y la probabilidad de que ocurran lesiones, y servirá de base para los siguientes capítulos del estudio.

```
RangeIndex: 908 entries, 0 to 907
Data columns (total 51 columns):
 #   Column                 Non-Null Count  Dtype
---  ------                 --------------  -----
 0   Indice                 908 non-null    int64
 1   Año                    908 non-null    int64
 2   Competicion            908 non-null    object
 3   Partido                908 non-null    object
 4   Fase                   905 non-null    object
 5   Fecha                  908 non-null    datetime64[ns]
 6   Mes                    908 non-null    int64
 7   Trimestre              908 non-null    int64
 8   Ciudad                 907 non-null    object
 9   Pais                   908 non-null    object
 10  Altura                 907 non-null    float64
 11  Temperatura            577 non-null    float64
 12  Distancia visitante    907 non-null    float64
 13  Distancia local        908 non-null    int64
 14  Lesion local           907 non-null    float64
 15  Lesion visitante       907 non-null    float64
 16  Lesion total           907 non-null    float64
 17  Diagnostico local 1    183 non-null    float64
 18  Localizacion local 1   183 non-null    float64
 19  Minuto local 1         193 non-null    float64
 20  Diagnostico visita 1   141 non-null    float64
 21  Localizacion visita 1  141 non-null    float64
 22  Minuto visita 1        128 non-null    object
 23  Diagnostico local 2    19 non-null     float64
 24  Localizacion local 2   19 non-null     float64
 25  Minuto local 2         19 non-null     float64
 26  Diagnostico visita 2   15 non-null     float64
 27  Localizacion visita 2  15 non-null     float64
 28  Minuto visita 2        15 non-null     float64
 29  Diagnostico local 3    2 non-null      float64
 30  Localizacion local 3   2 non-null      float64
 31  Minuto local 3         2 non-null      float64
 32  Diagnostico visita 3   2 non-null      float64
 33  Localizacion visita 3  2 non-null      float64
 34  Minuto visita 3        2 non-null      float64
 35  TEAM1                  908 non-null    object
 36  TEAM2                  908 non-null    object
 37  TEAM1_NAME             908 non-null    object
 38  TEAM1_COUNTRY          908 non-null    object
 39  TEAM2_NAME             908 non-null    object
 40  TEAM2_COUNTRY          908 non-null    object
 41  DIAS_TRANSCURRIDOS     908 non-null    int64
 42  PAIS_PARTIDO           908 non-null    object
 43  CIUDAD_PARTIDO         907 non-null    object
 44  FASE_PARTIDO           908 non-null    object
 45  IDA_VUELTA             131 non-null    object
 46  FASE_AGRUPADA          908 non-null    object
 47  TEAM1_NAME_NORM        908 non-null    object
 48  TEAM2_NAME_NORM        908 non-null    object
 49  TEAM1_NAME_FINAL       908 non-null    object
 50  TEAM2_NAME_FINAL       908 non-null    object
dtypes: datetime64[ns](1), float64(23), int64(6), object(21)
memory usage: 361.9+ KB
```

Figura 5: Columnas del dataset limpiado

5.3 ANÁLISIS DESCRIPTIVO DE LAS VARIABLES

5.3.1 COMPETICIÓN

La variable COMPETICION clasifica cada partido según el torneo continental al que pertenece: la Copa CONMEBOL Libertadores o la Copa CONMEBOL Sudamericana. Esta distinción es fundamental, ya que ambas competiciones presentan diferencias en el nivel competitivo, la distribución geográfica de los equipos y el calendario. La Copa Libertadores suele congregar a los equipos mejor posicionados en sus respectivas ligas nacionales, mientras que la Sudamericana acoge a equipos de rendimiento medio-alto o que han sido eliminados en fases preliminares de la Libertadores.

Además, Libertadores cuenta con 155 partidos por temporada [35], mientras que Sudamericana cuenta con 157 [36], lo que implicaría 465 y 471 partidos a lo largo de tres temporadas. Observando la Figura 6, se puede concluir que dentro de nuestro dataset nos faltan los datos de 5 partidos de Libertadores y 23 partidos de Sudamericana.

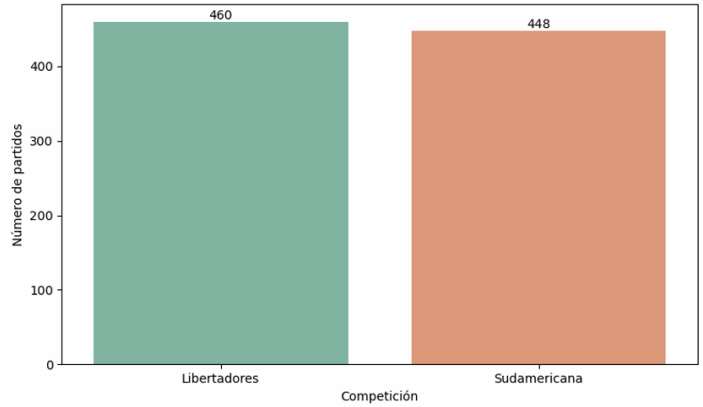

Figura 6: Número de partidos por competición

Como se ha mencionado en el punto 5.1.2, la Copa Libertadores tiene siempre 47 equipos participantes mientras que la Sudamericana varía y oscila alrededor de 56. Esto mismo se puede comprobar en la Figura 7.

	Año	Competicion	N_EQUIPOS_UNICOS
0	2022	Libertadores	47
1	2022	Sudamericana	57
2	2023	Libertadores	47
3	2023	Sudamericana	55
4	2024	Libertadores	47
5	2024	Sudamericana	53

Figura 7: Numero de partidos por competición y año

5.3.2 FASE

La variable FASE_PARTIDO indica la fase concreta en la que se disputa cada encuentro. Esta información resulta esencial para evaluar cómo varía la incidencia de lesiones en función del nivel de exigencia competitivo, ya que no es lo mismo un partido de fase clasificatoria que uno de eliminación directa.

Las categorías incluidas en esta variable son: Primera / Segunda / Tercera fase (rondas preliminares de clasificación), fase de grupos: partidos de en formato ida-vuelta por grupos y playoffs, octavos de final, cuartos de final, semifinal y final. El funcionamiento detallado del número de equipos y partidos en cada fase se encuentra en el punto 5.1.2.

Estos mismos datos se reflejan en la Figura 8, donde efectivamente, el dataset no incluye datos de semifinales ni final de la Copa Libertadores 2024, dando lugar a los 5 partidos ausentes para libertadores. En el caso de la Copa Sudamericana, se observa que para la temporada 2024, faltan 19 de los 157 partidos por la ausencia de 13 partidos de playoffs, uno de cuartos y las semifinales y final [37].

Además, en la Figura 8 se observa que en 2022 no hubo playoffs para Sudamericana y cuenta con una primera fase mucho más extensa. Esto se debe a que, en 2023, la CONMEBOL introdujo importantes modificaciones en el formato de la Copa Sudamericana. Por un lado, la primera fase pasó a disputarse mediante partidos únicos (eliminatorias directas entre equipos del mismo país) sustituyendo al formato tradicional de partidos ida-vuelta y se creó la fase de playoffs. Fase que enfrenta los 8 segundos clasificados de la fase de grupos de la

Sudamericana y los 8 terceros de grupo provenientes de la Libertadores, añadiendo así 8 nuevos cruces eliminatorios al calendario del torneo cambios que, según la CONMEBOL, generan dinamismo y atractivo mediático para la competición [38].

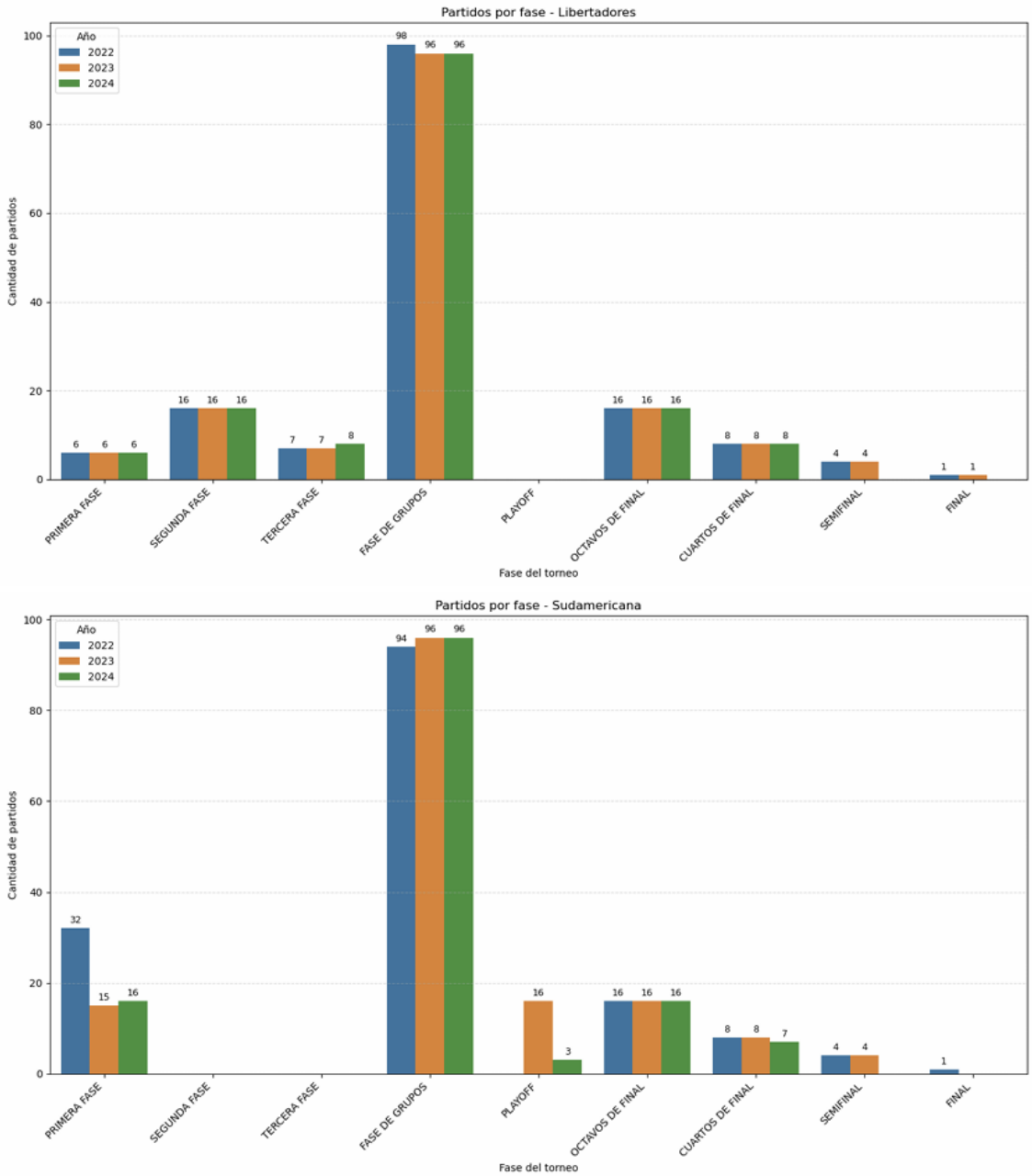

Figura 8: Partidos por fase y año, separado por competición

5.3.3 FECHA

5.3.3.1 Año

La variable AÑO indica el año calendario en que se disputó cada partido registrado en el dataset. El rango temporal abarca desde 2022 hasta 2024, lo que permite analizar posibles tendencias o cambios interanuales en la incidencia de lesiones. La reducción en el número de partidos disponibles para el año 2024 observada en la Figura 9 se debe a la falta de datos correspondientes a las fases finales de ambas competiciones, como se ha explicado previamente en el apartado 5.3.2.

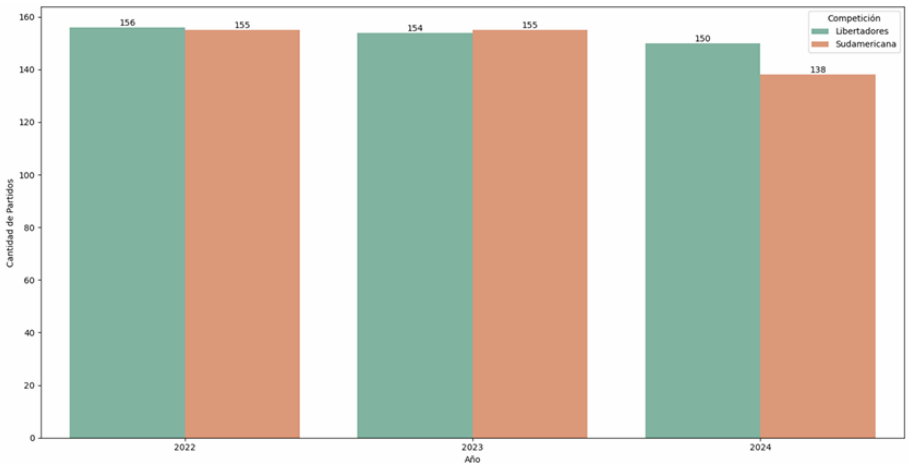

Figura 9: Número de partidos por año y competición

5.3.3.2 Mes y trimestre

La variable MES (rango de 2 a 11) permite analizar la distribución temporal de los partidos a lo largo del año. En la Figura 10 se observa que la mayoría de los encuentros se concentran en los meses de abril (249 partidos) y mayo (252 partidos), seguidos de marzo y agosto, con 104 y 99 partidos respectivamente. Esta distribución evidencia que la actividad futbolística en las competiciones latinoamericanas tiende a concentrarse en el primer semestre del año, en concordancia con los calendarios establecidos por la CONMEBOL para las fases de grupos y los inicios de las eliminatorias, con el 75% de los partidos celebrados antes de junio.

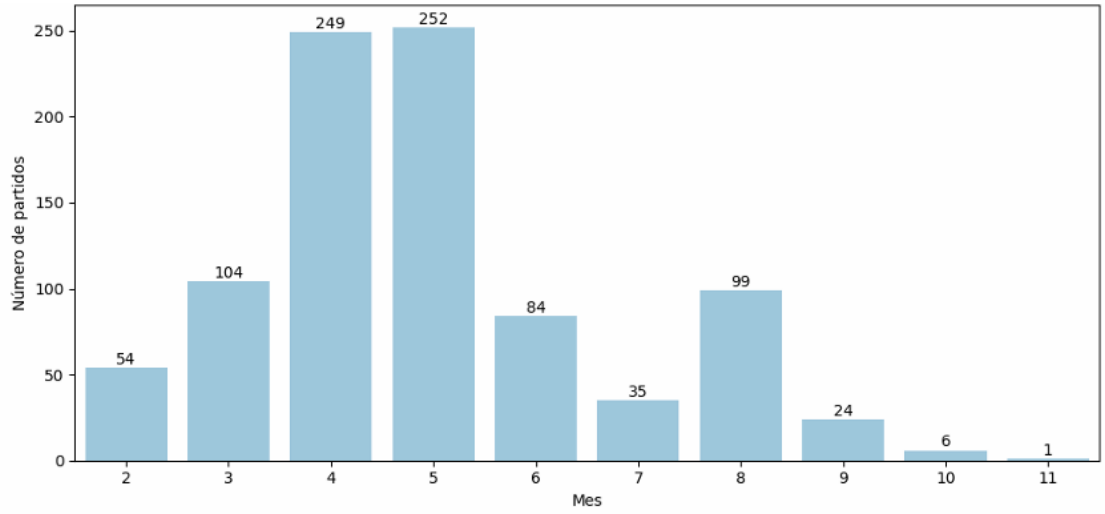

Figura 10: Número de partidos por mes

Sin embargo, en agosto de 2023 se observa un aumento atípico de encuentros, debido al retraso de los octavos de final de junio a agosto y la coincidencia entre octavos y cuartos, todo impulsado por la incorporación de la nueva ronda de playoffs en la Copa Sudamericana. Esto se observa comparando la Figura 11 con la Figura 12 y la Figura 13, donde se ve como la nueva fase (en verde) ha retrasado el comienzo de octavos de final (en rojo). Estas tendencias no solo reflejan el diseño de calendario competitivo, sino que también son relevantes para entender el impacto de la acumulación de carga física sobre el riesgo de lesión en determinadas épocas del año.

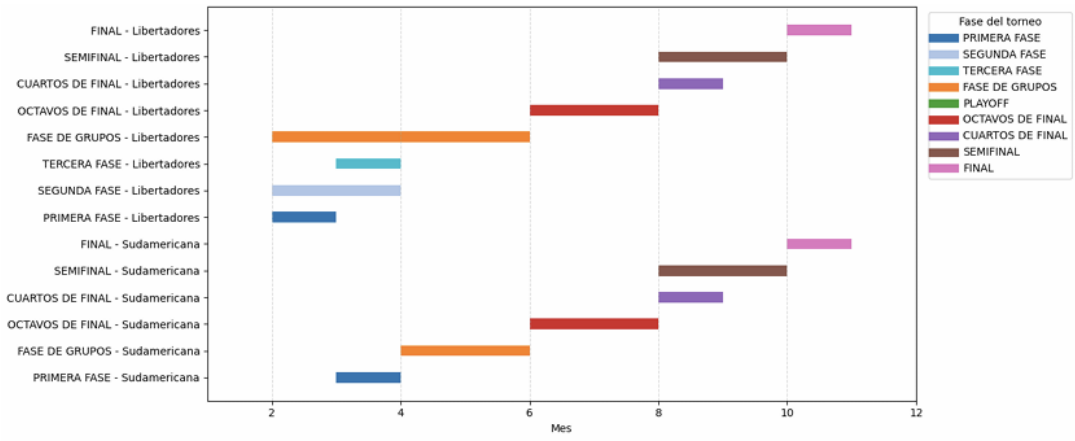

Figura 11: Calendario fases - 2022

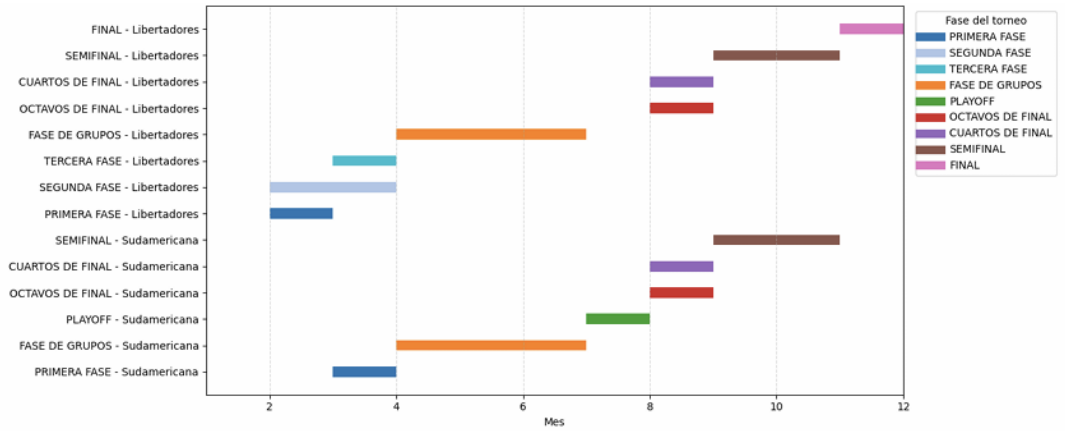

Figura 12: Calendario fases - 2023

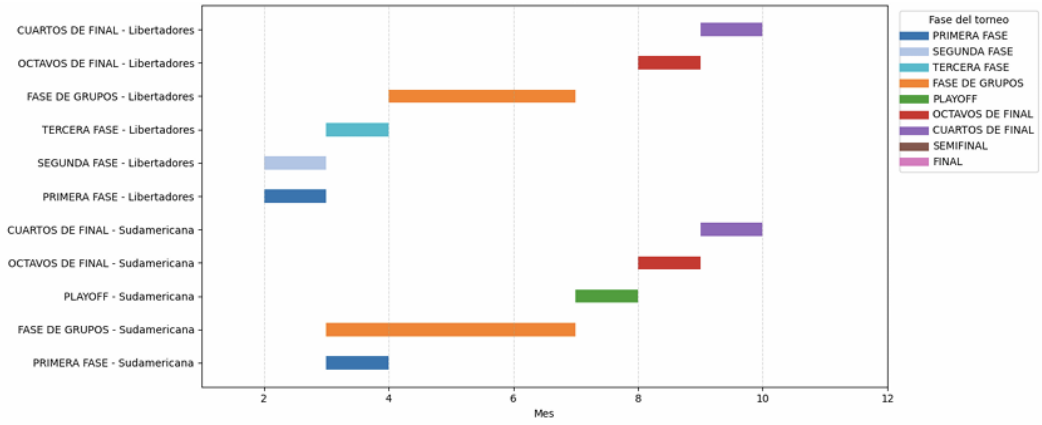

Figura 13: calendario fases - 2024

5.3.4 VARIABLES GEOGRÁFICAS

5.3.4.1 Altitud

La variable **ALTURA**, expresada en metros sobre el nivel del mar, representa la elevación de la ciudad donde se disputa cada encuentro. Su análisis revela una distribución altamente asimétrica, con una media de 725 metros y una mediana significativamente inferior (160 m), lo que sugiere una gran concentración de partidos en zonas de baja altitud. De hecho, el 25 % de los partidos se juega por debajo de los 25 metros, y el 75 % por debajo de los 900 metros.

Esta tendencia queda reflejada en el histograma y el boxplot de la Figura 14, donde se observa una clara acumulación de eventos cerca del nivel del mar, como pueden ser las

ciudades de Buenos Aires, Río de Janeiro o Montevideo y una larga cola hacia valores extremos. Entre los valores atípicos pueden destacar partidos jugados en ciudades como La Paz (3 640 m), Quito (2 850 m) o Cuzco (3 399 m).

Estudios señalan que altitudes superiores a 2500 m afectan la oxigenación, provocan acidosis láctica más temprana y fatiga muscular, aumentando la susceptibilidad a lesiones musculares [39]. Ejemplos recientes como el del joven jugador del Palmeiras, Estêvão, que a pesar de tener 18 años y un físico resistente, tuvo que ser retirado en camilla tras sufrir vómitos y falta de oxígeno en La Paz, refuerzan esta afirmación [40].

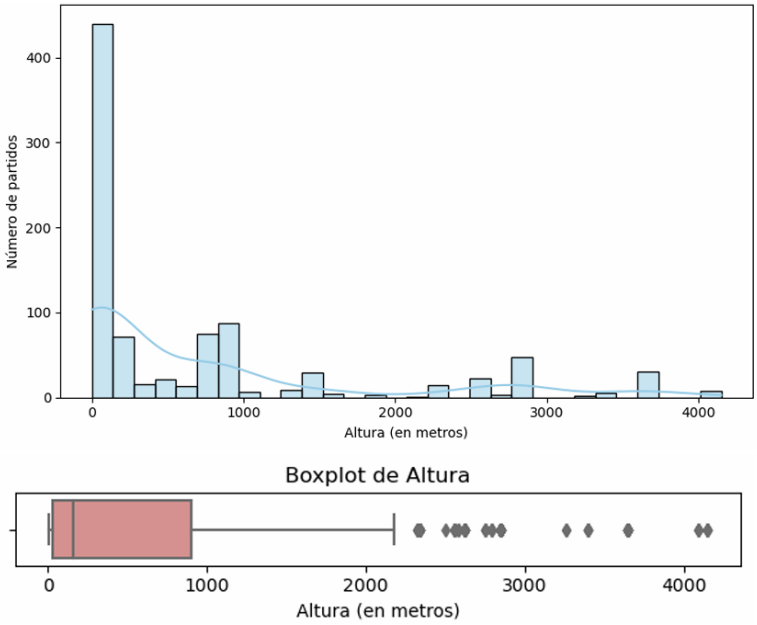

Figura 14: Distribución de altura de las ciudades en las que se disputan los partidos

5.3.4.2 Temperatura

La variable **TEMPERATURA**, registrada en grados Celsius, refleja el clima estimado en el momento de cada partido y solo es conocida para 577 de los 908 partidos. En la Figura 15 se ve que su distribución es ligeramente asimétrica hacia la derecha, con una media de 19,07 °C y una mediana casi idéntica (19 °C), lo que indica una tendencia clara hacia temperaturas moderadas. El rango completo oscila entre 7 °C y 36 °C, lo que evidencia la

enorme variedad térmica del contexto latinoamericano, desde zonas frías hasta regiones tropicales.

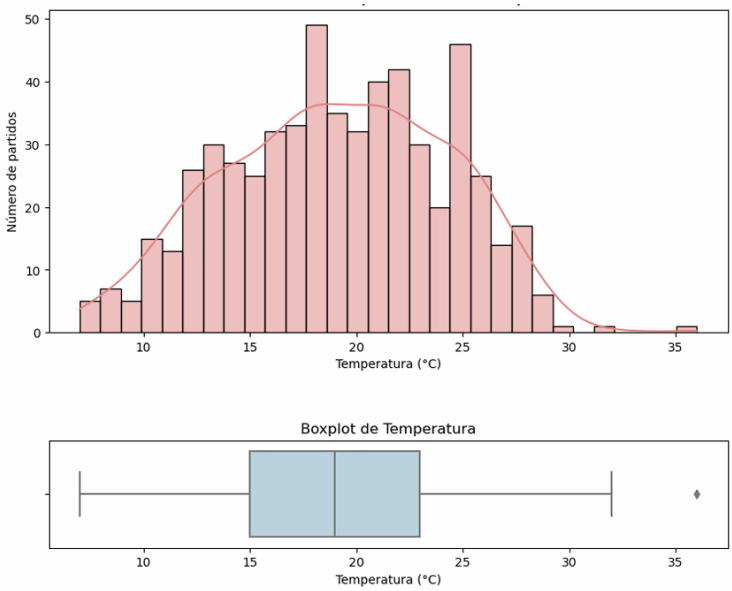

Figura 15: Distribución de la temperatura en los partidos

La mayor parte de los partidos, según se observa en la Figura 15 se disputan entre 15 °C y 25 °C, rango considerado óptimo para el rendimiento físico. Sin embargo, el histograma muestra **outliers** por **encima de los 30 °C,** lo cual puede representar un riesgo añadido. Según un estudio publicado en The Journal of Strength and Conditioning Research, temperaturas elevadas están asociadas con una reducción del ritmo de juego (pacing) y un mayor esfuerzo fisiológico, factores que pueden comprometer el rendimiento y aumentar la probabilidad de lesiones relacionadas con la fatiga o el calor [41].

5.3.5 LOCALIZACIÓN

5.3.5.1 País y Ciudad

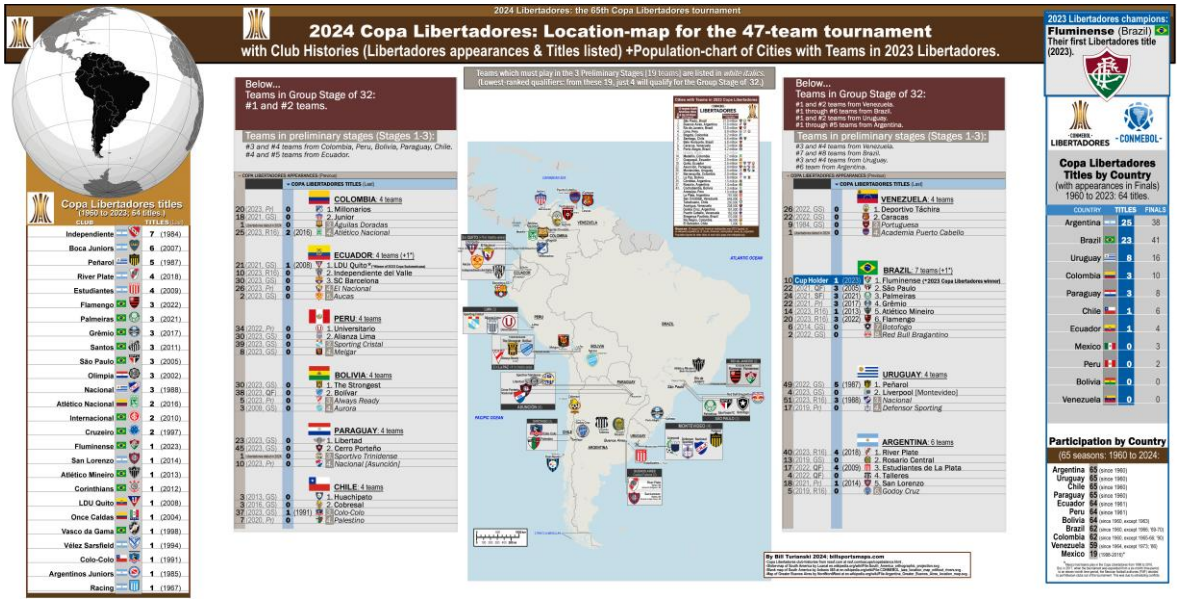

Figura 16: Mapa con los países participantes en la copa libertadores [42]

El país y la ciudad donde se celebra el partido son variables clave para entender el contexto ambiental, logístico y competitivo en el que se desarrollan las competiciones latinoamericanas. Estas variables permiten estudiar correlaciones con factores como la altitud, la temperatura, los desplazamientos o la presión institucional sobre los equipos. En la Figura 16, se puede ver un mapa de todos los países con sus estadios participantes, lo cual nos ayuda a hacer una idea tanto de temperaturas y altitud, como de distancias recorridas según el país anfitrión.

Como muestra la Figura 17, **Brasil (218 partidos)** y **Argentina (154 partidos)** lideran el número de encuentros disputados en el periodo 2022–2024. Esta distribución es coherente con el número de equipos que representan a ambos países, así como con la infraestructura disponible para albergar encuentros internacionales.

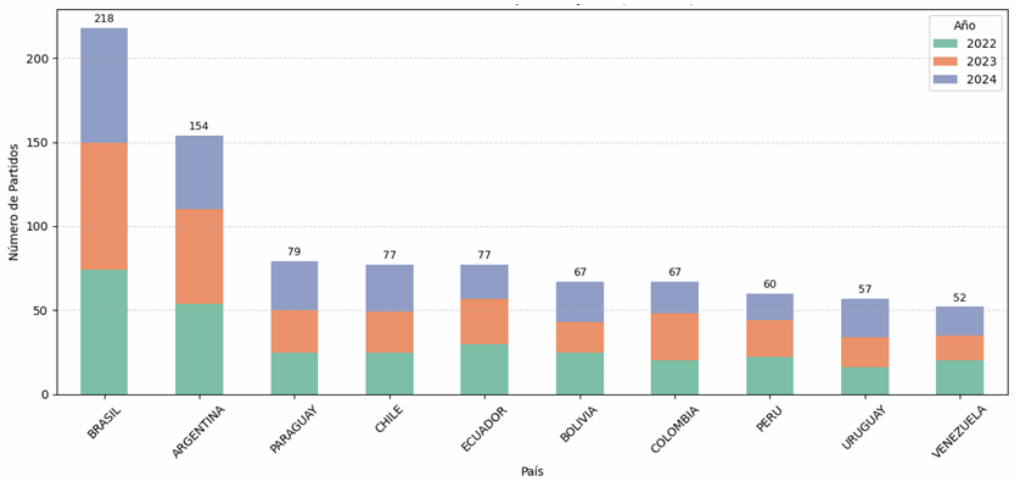

Figura 17: Partidos por país

En el análisis por ciudad Figura 18, Buenos Aires destaca con diferencia como la sede más frecuente, acumulando 92 partidos. Este dato refleja su condición histórica como capital futbolística regional: **Buenos Aires es la ciudad con más estadios de fútbol profesionales del mundo, con al menos 18 recintos en uso**, lo que facilita la organización simultánea de partidos y el uso de múltiples sedes en una misma temporada [43].

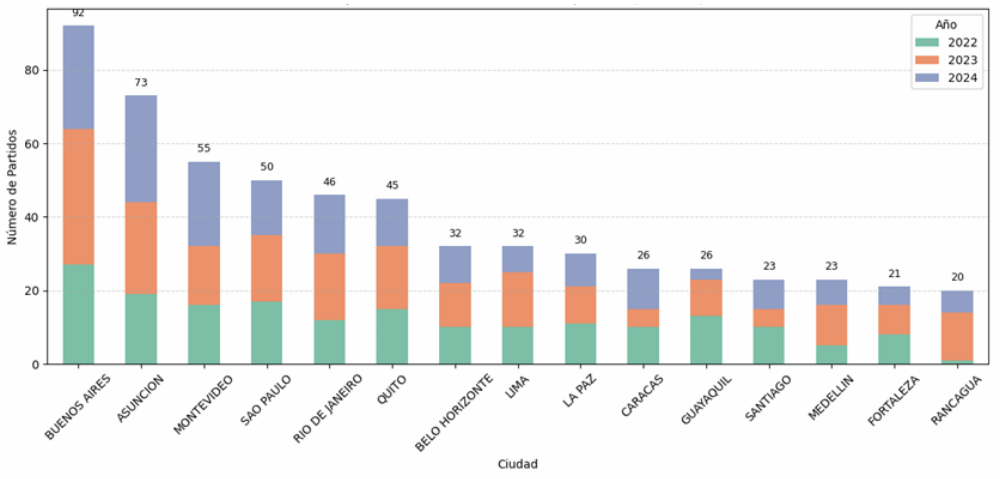

Figura 18: Número de partidos por ciudad

Otros núcleos urbanos con alta recurrencia son **Asunción**, **Montevideo**, **São Paulo**, **Quito** y **La Paz**, lo pone en relieve la relevancia de la geografía competitiva que varía desde centros urbanos a nivel del mar hasta ciudades a gran altitud como La Paz o Quito. Diferencias que pueden condicionar el esfuerzo físico, la recuperación y, en consecuencia, el riesgo de lesión.

5.3.5.2 Distancia visitante

La variable Distancia visitante cuantifica los kilómetros recorridos por los equipos en condición de visitante para disputar cada partido. El análisis revela una **amplia dispersión geográfica** en las competiciones latinoamericanas, reflejo de las grandes extensiones territoriales de los países miembros de la CONMEBOL. Aunque el continente solo agrupa a diez naciones, algunas como Brasil, Argentina o Colombia abarcan vastas áreas, y otras, como Venezuela o Bolivia, están alejadas de los principales núcleos futbolísticos.

Desde el punto de vista estadístico, el boxplot de la Figura 19, la distancia media recorrida por partido es de 2.400 km, con una desviación estándar de 1.557 km, lo que evidencia una gran heterogeneidad. El 50% de los partidos implican viajes superiores a los 2.200 km, **y el 25% superan los 4.000 km,** un dato significativo cuando se considera el alto ritmo competitivo y el escaso margen de descanso entre encuentros. Se ha demostrado que este tipo de desplazamientos prolongados, en los que pueden variar zonas horarias, altitud y clima, afectan negativamente al rendimiento físico y aumentan el riesgo de lesiones musculares o por fatiga [44].

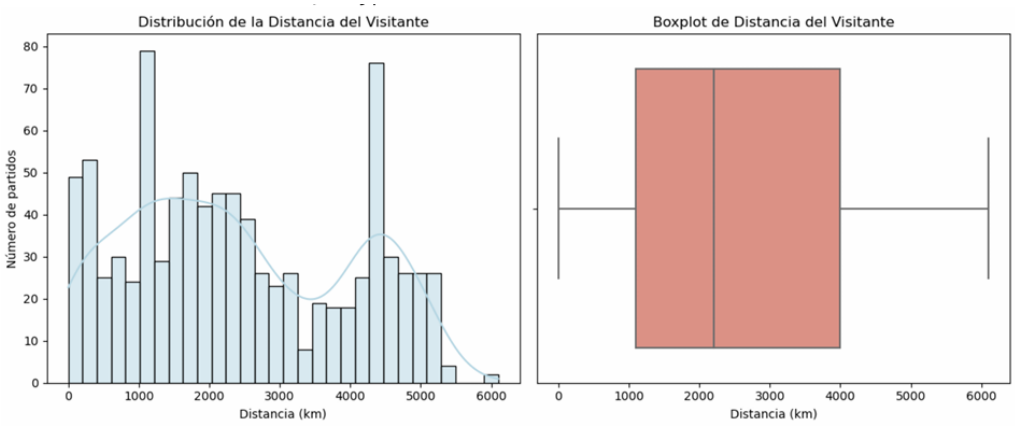

Figura 19: Estadística de la distancia recorrida por el visitante

La distancia recorrida cobra aún más relevancia al analizar los equipos que han alcanzado las finales, en este caso de la Copa Sudamericana:

- En **2022**, Independiente del Valle (Ecuador) y Sao Paulo (Brasil) disputaron la final en Córdoba, Argentina, recorriendo aproximadamente 4.100 km y 2.300 km respectivamente [45].
- En **2023**, los finalistas fueron LDU Quito (Ecuador) y Fortaleza (Brasil), quienes jugaron en Maldonado, Uruguay, recorriendo entre 4.000 y 4.300 km, un 15-20% más que la mayoría de los equipos con solo un partido [46].

En ambos casos, como se ve en la Figura 20, los equipos finalistas están entre los equipos con más distancia recorrida para la temporada, indicando también un mayor cansancio acumulado y por lo tanto, más probabilidad de padecer una lesión.

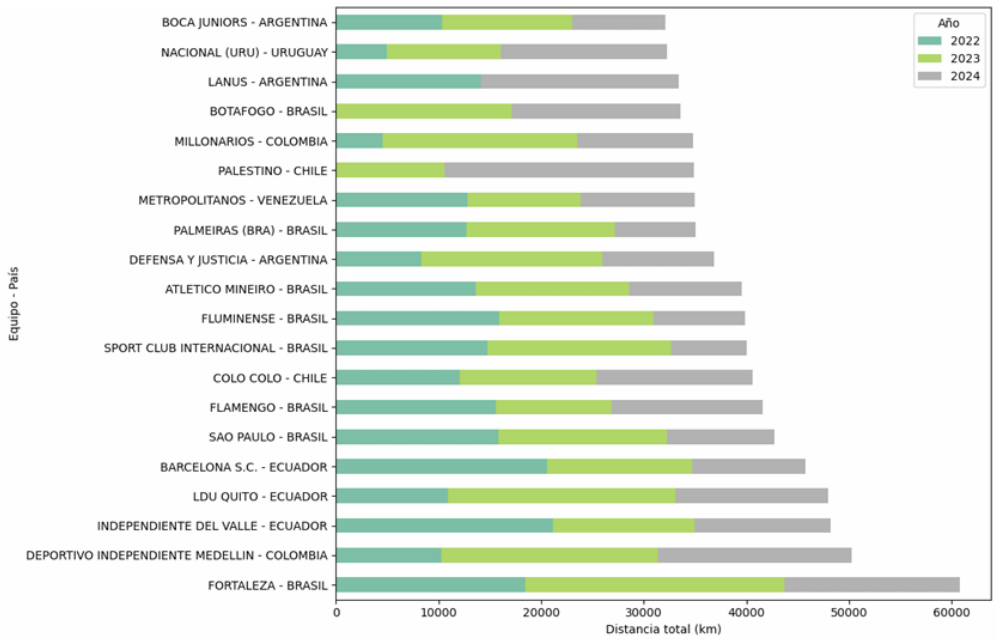

Figura 20: Top 20 equipos con más distancia recorrida

Por otro lado, el desglose por países en la Figura 21 revela patrones interesantes. **Venezuela, Colombia y Ecuador** encabezan la lista de países cuyos equipos recorren mayores distancias promedio por partido.

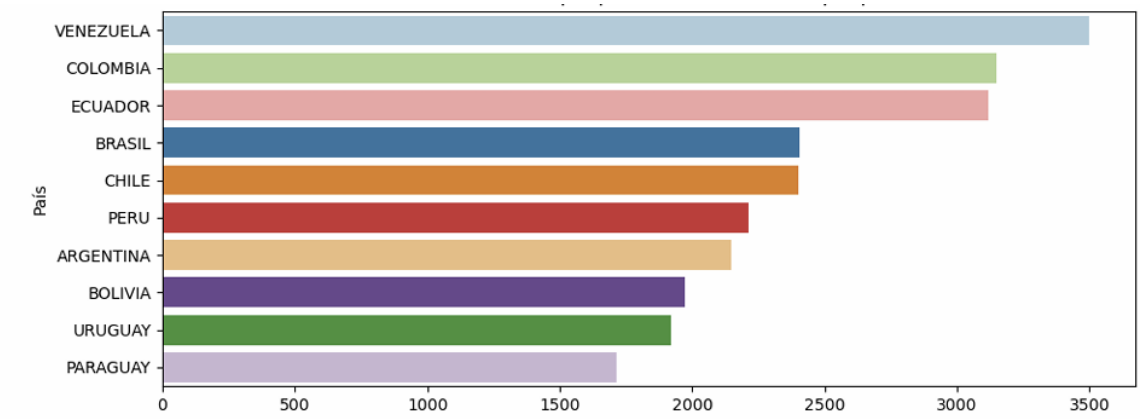

Figura 21: Distancia media recorrida por cada país como equipo visitante

En el caso de Venezuela, que lidera en distancia media por país, se debe en gran parte a su localización periférica respecto al núcleo futbolístico latinoamericano, lo que implica vuelos largos a la mayoría de los países. Además, muchos de sus equipos compiten en fases previas, lo que implica viajes adicionales antes de ser eliminados. En el caso de Ecuador, esta media elevada se debe tanto a la localización como al buen rendimiento de sus clubes, que suelen avanzar a fases finales.

5.3.6 EQUIPOS

5.3.6.1 Nombre y país del equipo

En la Figura 22, se muestran los equipos representados y el país al que pertenecen, lo que resalta que Brasil y Argentina tienen una presencia superior al resto. Esto se debe a que, gracias a los cupos de la CONMEBOL, basados en coeficientes y rendimiento histórico, se les asignan más equipos. En la **Copa Libertadores** (2025) Brasil dispone de **7 plazas directas + 1 por campeón vigente**, mientras que Argentina tiene **6 plazas + 1 por campeón** mientras que los ocho países restantes tienen 4 plazas cada uno [35].

En la **Copa Sudamericana** (2024), tanto Brasil como Argentina tienen **6 plazas cada uno**, frente a las 4 que corresponden a cada una de las federaciones restantes [36].

	PAIS	NUM_EQUIPOS	EQUIPOS
0	ARGENTINA	22	ARGENTINOS JUNIORS, BANFIELD, BELGRANO, BOCA JUNIORS, C.A. TALLERES CBA., COLON, DEFENSA Y JUSTICIA, ESTUDIANTES DE LA PLATA, GIMNASIA LA PLATA, GODOY CRUZ, HURACAN, INDEPENDIENTE (ARG), LANUS, NEWELL'S OLD BOYS, PATRONATO, RACING (ARG), RIVER PLATE (ARG), ROSARIO CENTRAL, SAN LORENZO, TIGRE, UNION, VELEZ
2	BRASIL	19	AMERICA (BRA), ATLETICO CLUBE GOIANIENSE, ATLETICO MINEIRO, ATLETICO PARANAENSE, BOTAFOGO, CEARA SPORTING CLUB, CRUZEIRO, CUIABA EC, FLAMENGO, FLUMINENSE, FORTALEZA, GOIAS, GREMIO, PALMEIRAS (BRA), RED BULL BRAGANTINO, SANTOS, SAO PAULO, SPORT CLUB CORINTHIANS PAULISTA, SPORT CLUB INTERNACIONAL
3	CHILE	14	ANTOFAGASTA, AUDAX ITALIANO, COBRESAL, COLO COLO, COQUIMBO UNIDO, CURICO UNIDO, EVERTON, HUACHIPATO, MAGALLANES, NUBLENSE, PALESTINO, UNION ESPANOLA, UNION LA CALERA, UNIVERSIDAD CATOLICA (CHI)
4	COLOMBIA	14	AGUILAS DORADAS, ALIANZA (COL), AMERICA (COL), ATLETICO NACIONAL, DEPORTES TOLIMA, DEPORTIVO CALI, DEPORTIVO INDEPENDIENTE MEDELLIN, DEPORTIVO PEREIRA, EQUIDAD, INDEPENDIENTE SANTA FE, JUNIOR, LA EQUIDAD, MILLONARIOS, PALMEIRAS (COL)
1	BOLIVIA	13	ALWAYS READY, ATLETICO PALMA FLOR, AURORA, BLOOMING, BOLIVAR, GUABIRA, INDEPENDIENTE (BOL), JORGE WILSTERMANN, ORIENTE PETROLERO, REAL TOMAYAPO, ROYAL PARI, THE STRONGEST, UNIVERSITARIO (BOL)
7	PERU	13	ALIANZA LIMA, ASOCIACION DEPORTIVA TARMA, AYACUCHO FC, BINACIONAL, CIENCIANO, DEPORTIVO GARCILASO, F.B.C. MELGAR, SPORT BOYS, SPORT HUANCAYO, SPORTING CRISTAL, UNIVERSIDAD CATOLICA (PER), UNIVERSIDAD CESAR VALLEJO, UNIVERSITARIO (PER)
8	URUGUAY	13	BOSTON RIVER, CERRO LARGO, DANUBIO, DEFENSOR SPORTING, DEPORTIVO MALDONADO, LIVERPOOL, MONTEVIDEO CITY TORQUE, MONTEVIDEO WANDERERS, NACIONAL (URU), PENAROL, PLAZA COLONIA, RACING (URU), RIVER PLATE (URU)
9	VENEZUELA	13	ACADEMIA PUERTO CABELLO, C.D. HERMANOS COLMENAREZ, CARABOBO, CARACAS, DEPORTIVO LA GUAIRA F.C., DEPORTIVO LARA, DEPORTIVO TACHIRA, ESTUDIANTES DE MERIDA, METROPOLITANOS, MONAGAS, PORTUGUESA, RAYO ZULIANO, ZAMORA
5	ECUADOR	12	A.D.NUEVE DE OCTUBRE, AUCAS, BARCELONA S.C., C.S.EMELEC, DELFIN, DEPORTIVO CUENCA, EL NACIONAL, INDEPENDIENTE DEL VALLE, LDU QUITO, MUSHUC RUNA, TECNICO UNIVERSITARIO, UNIVERSIDAD CATOLICA (ECU)
6	PARAGUAY	12	CERRO PORTENO, GENERAL CABALLERO JLM, GUAIRENA FUTBOL CLUB, GUARANI, LIBERTAD, NACIONAL (PAR), OLIMPIA, SOL DE AMERICA, SPORTIVO AMELIANO, SPORTIVO LUQUENO, SPORTIVO TRINIDENSE, TACUARY

Figura 22: Tabla de equipos y país

5.3.7 VARIABLES TEMPORALES

5.3.7.1 Días transcurridos y comienzos de temporada

La variable DÍAS_TRANSCURRIDOS cuantifica cuántos días han pasado desde el 1 de enero hasta la fecha en la que se disputa cada partido. Esta métrica permite estimar en qué momento del año se encuentra cada encuentro, ofreciendo una aproximación a la posible carga física acumulada de los jugadores.

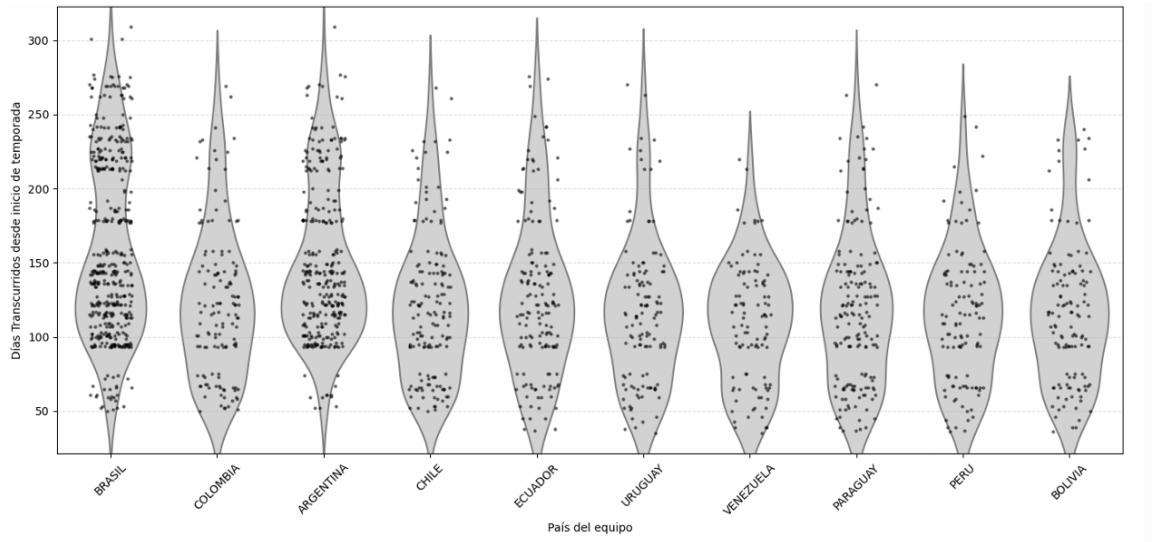

Figura 23: Distribución de los días transcurridos por país del equipo

El gráfico de violín de la Figura 23 permite visualizar la densidad y dispersión por país, revelando los siguientes patrones:

- **Brasil y Argentina** presentan distribuciones más extendidas, con valores altos al final del año. Esto se debe a que sus clubes más representativos acceden directamente a la fase de grupos y llegan a fases finales, por lo que su participación se prolonga hasta los últimos meses de competición. Además, sus campeonatos nacionales comienzan más tarde, el 13 de abril para el Campeonato Brasileño Serie A en 2024 [47] y 12 de mayo para la Liga Profesional Argentina en 2024 [48]. Esto puede permitir a los jugadores llegar en mejores condiciones a las fases más competitivas.

- **Países como Paraguay o Perú**, cuyas ligas arrancan en enero, presentan distribuciones más comprimidas en los primeros tramos del año.

- La dualidad de torneos nacionales en Colombia (Apertura y Finalización) o Ecuador (con etapas divididas) también genera una variedad de fechas posibles en las que los clubes están activos.

5.3.8 VARIABLES DE LESIÓN

5.3.8.1 Lesión

Tenemos las variables lesión local y lesión visitante que como su nombre indican el número de lesiones en el equipo local o visitante y luego la variable lesión total que agrupa ambas.

5.3.8.2 Diagnóstico y localización

Columnas formadas por diferentes códigos, cada uno correspondiente a un tipo de diagnóstico o localización de la lesión (consultar ANEXO II). Se han representado en la Figura 24 y la Figura 25 el número de lesiones por tipo de diagnóstico y localización, respectivamente, según el año y en variables absolutas y relativas. Las lesiones musculares como contracturas y desgarros en zonas inferiores del cuerpo como los músculos isquiotibiales y el resto de la pierna son las más comunes a lo largo de todas las temporadas. Además, los diagnósticos se han agrupado en 13 grupos y las localizaciones en 5 para facilitar las visualizaciones y su uso en los modelos en los capítulos siguientes, quedando los resultados reflejados en la Figura 26 y la Figura 27.

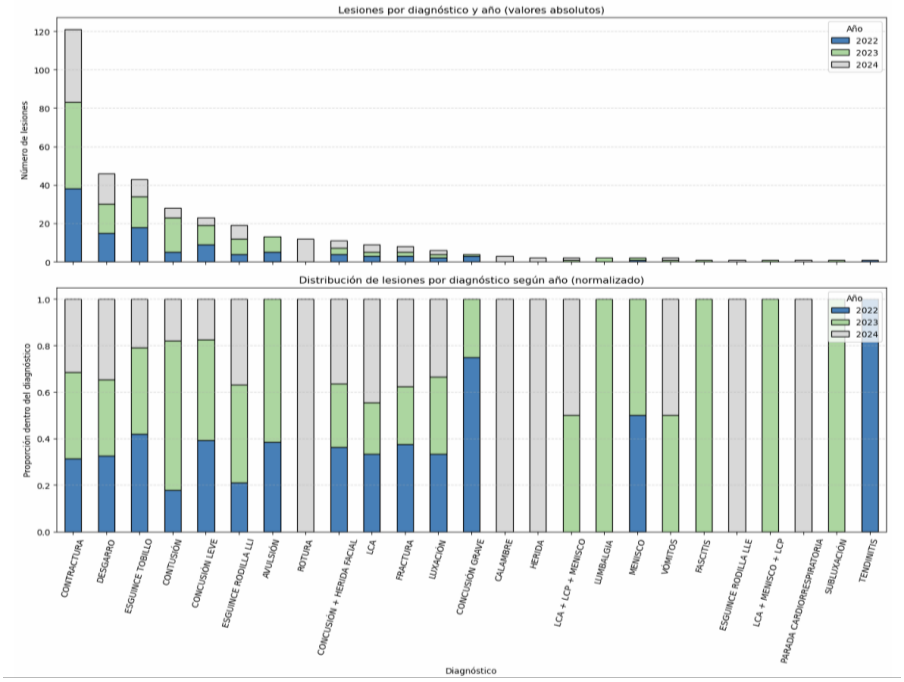

Figura 24: Distribución de número de lesiones por diagnóstico

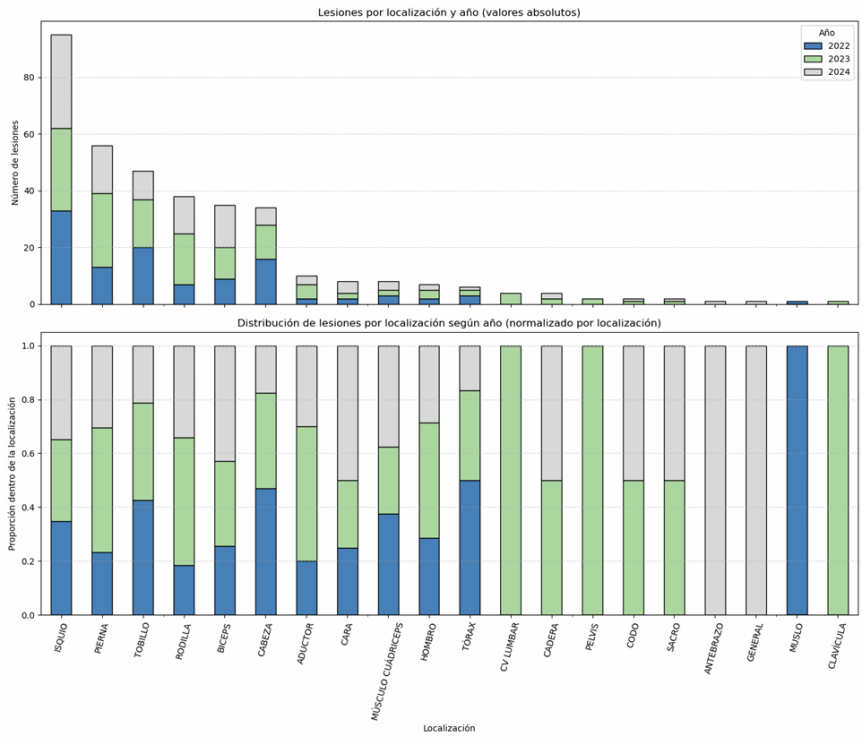

Figura 25: Distribución de localización es de las lesiones

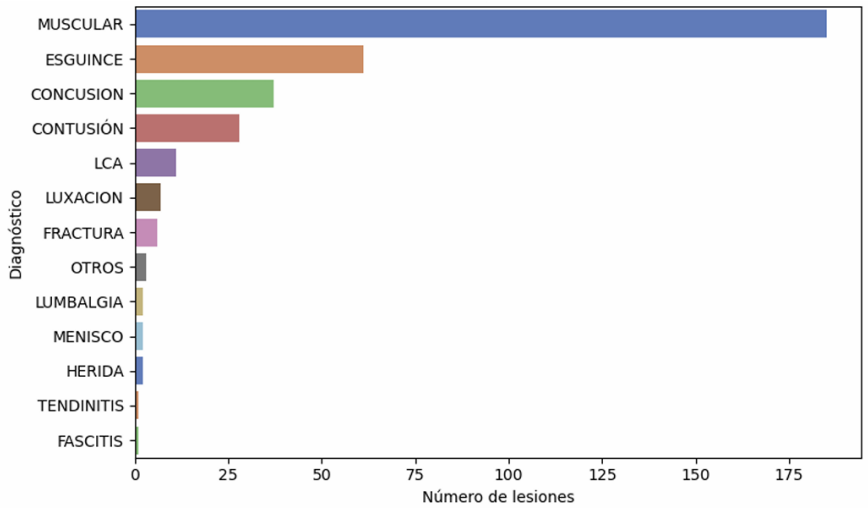

Figura 26: Número de lesiones por tipo de diagnóstico agrupado

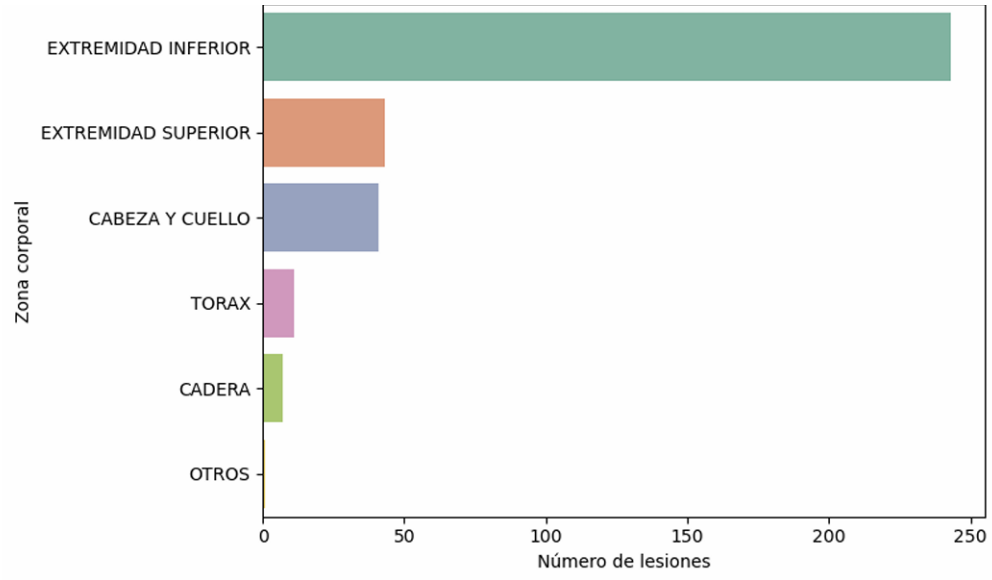

Figura 27: Número de lesiones según la zona corporal

5.3.8.3 Minuto

De las 358 lesiones con minuto registrado, la media se sitúa en el minuto 46.8, lo que sugiere que la mayoría ocurren al inicio del segundo tiempo. La distribución presenta dos picos destacados: Entre los minutos 15 y 40, cuando el partido se intensifica y entre los minutos 65 y 85, coincidiendo con el cansancio muscular acumulado y mayor nivel competitivo.

Estos resultados se alinean con estudios científicos que indican un aumento del riesgo lesional a medida que avanza el encuentro. En particular, un estudio sobre futbolistas universitarios encontró que el riesgo de lesiones es mayor en el tercer y cuarto cuartos del partido (minutos 45–60 y 60–75), debido a la fatiga acumulada, deshidratación, menor fuerza muscular y alteraciones biomecánicas que afectan la toma de decisiones y la coordinación motora [49].

En cuanto a la localización (Figura 28), se observa una alta frecuencia de lesiones musculares (isquiotibiales, cuádriceps, aductores), especialmente en los tramos finales de cada mitad, lo que se puede deber a lesiones de sobrecarga muscular por fatiga o pérdida de eficiencia en la activación neuromuscular.

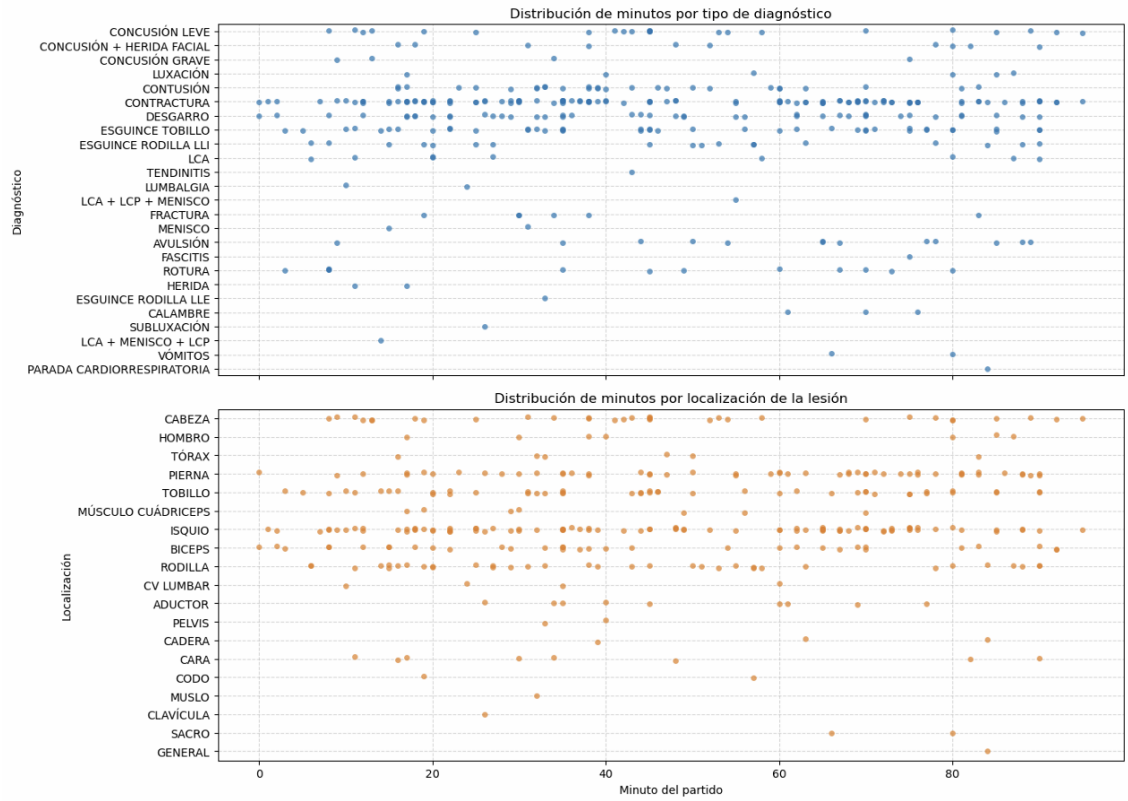

Figura 28: Distribución temporal por tipo de diagnóstico

5.4 ANÁLISIS EXPLORATORIO

El objetivo de esta sección es explorar visual y estadísticamente los factores que pueden influir en la aparición de lesiones durante los partidos de fútbol en competiciones CONMEBOL. A partir del análisis exploratorio y de las visualizaciones, se pretende generar hipótesis fundamentadas que sirvan como base para la siguiente fase del trabajo: la creación de un modelo predictivo de riesgo de lesión.

En particular, se busca estudiar cómo variables como la altura de la ciudad, la temperatura durante el partido, la distancia recorrida por los equipos, el momento de la temporada, o incluso el tipo de competición y la fase, pueden estar relacionadas con una mayor o menor incidencia de lesiones. Entre las preguntas que guían este análisis destacan:

- ¿Se producen más lesiones en ciertas fases del torneo o meses del año?
- ¿Afecta negativamente la altitud o el calor al número de lesiones?
- ¿Los equipos que viajan más sufren más lesiones?
- ¿Qué tipos de lesiones y localizaciones son más frecuentes?

Cabe destacar que este

Este análisis parte del dataset previamente limpiado y estructurado, cuyo procesamiento completo se recoge en el apartado 5.2. Las variables han sido estandarizadas y se han creado columnas derivadas que permiten un estudio más profundo del fenómeno de las lesiones.

5.4.1 CORRELACIONES INICIALES

El análisis de correlaciones de la Figura 29 muestra que no existe una relación lineal clara entre factores contextuales como la altitud, la temperatura o la distancia recorrida y la aparición de lesiones. Por ejemplo, la **altitud presenta una correlación negativa muy baja con las lesiones totales (-0.06), al igual que la temperatura (0.05) y la distancia recorrida (-0.04).** Estos valores sugieren que, si bien estas variables son relevantes desde el punto de vista fisiológico y logístico, su efecto no se manifiesta de forma directa ni lineal. El paso del tiempo en la temporada (medido como días transcurridos) también muestra correlaciones

casi nulas con el número de lesiones, lo que sugiere que factores como la carga acumulada no se reflejan automáticamente en esta variable, al menos en términos simples.

Estos resultados iniciales nos indican que no basta con observar relaciones directas, sino que debemos capturar interacciones de variables y posibles efectos específicos de contextos geográficos o competitivos para comprender mejor que variables explican el riesgo de lesión. Además, no bastará con aplicar modelos lineales simples en el punto 6.3, será necesario el uso de modelos que capturen relaciones más complejas para realizar una mejor predicción.

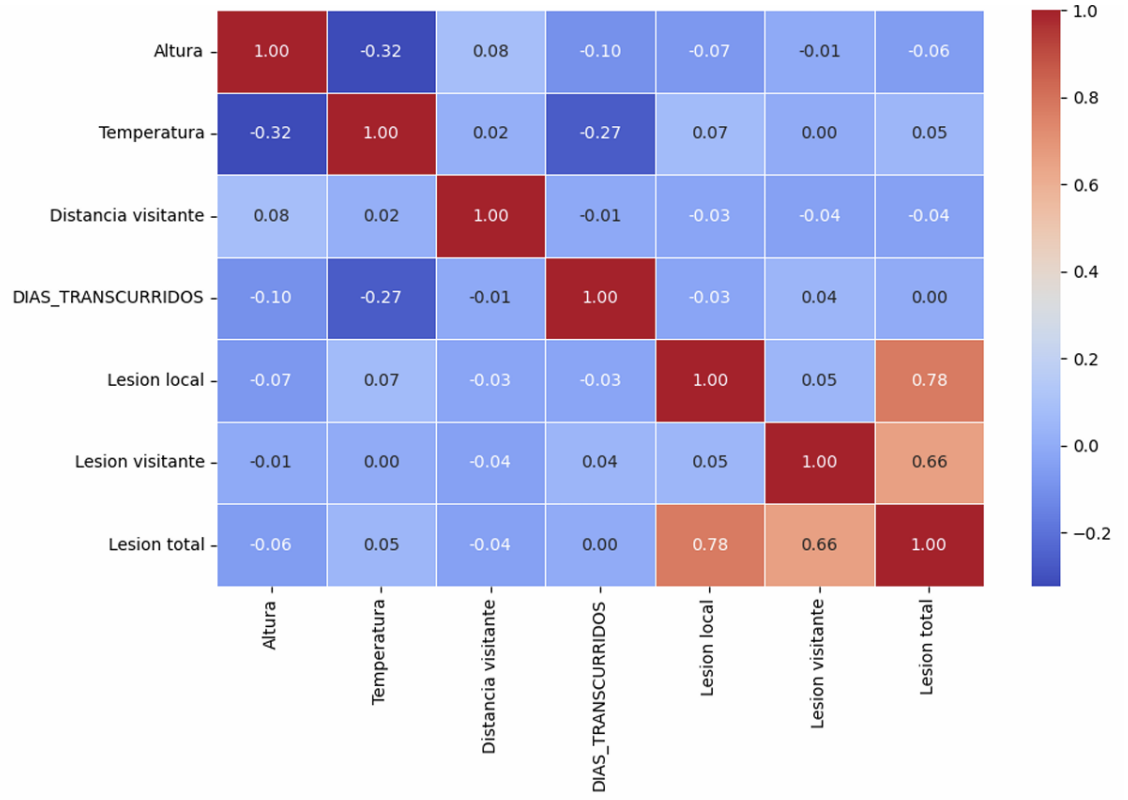

Figura 29: Matriz de correlaciones

5.4.2 COMPETICIÓN Y AÑO

Entre 2022 y 2023 se observa un aumento notable en el número absoluto de lesiones, especialmente en la Copa Sudamericana, aunque también en la Libertadores (Figura 30). Este repunte coincide con los cambios introducidos por CONMEBOL en el calendario de la Sudamericana en 2023, descritos en el apartado 5.3.2. Según la Conmebol, estos cambios

que reducen la fase de clasificación y añaden una ronda de play-offs, tienen el objetivo de **"aumentar el dinamismo y la exigencia"** para introducir mayor competitividad y aumentar el atractivo a la copa Sudamericana [38]. Este nuevo formato puede haber generado **mayor exigencia física** sobre los jugadores, especialmente en una primera temporada donde las rutinas de entrenamiento no se han adaptado a este nuevo calendario. Dado que muchos clubes comparten participación entre ambas competiciones, esta carga adicional puede impactar indirectamente en la Copa Libertadores.

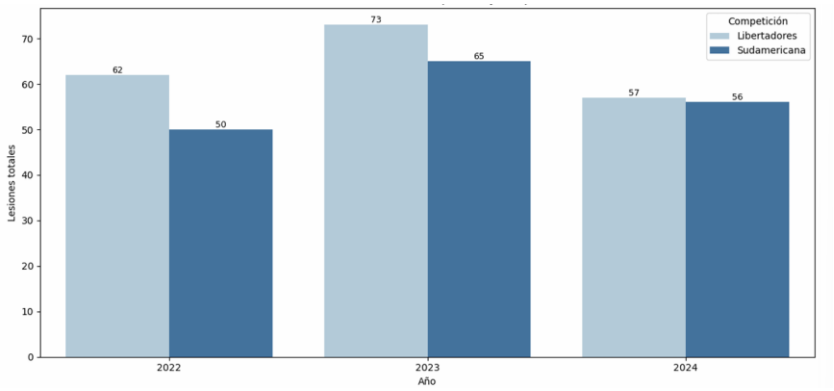

Figura 30: Número de lesiones por competición y año

Aunque en 2024 el número absoluto de lesiones parece reducirse, cabe destacar que como se mencionó en el apartado 5.3.1, no tenemos los datos de la temporada completa de 2024. Por ello, estudiamos porcentaje de partidos con al menos una lesión por año. En la Figura 31: % relativo de partidos con lesión por competiciónFigura 31 se observa que, tanto en Libertadores como Sudamericana, la incidencia de lesiones de 2024 es superior a la de 2022, aunque con una ligera moderación respecto a 2023. Esto sugiere que los equipos podrían estar empezando a adaptarse al nuevo formato competitivo, ajustando progresivamente sus estrategias de carga, rotación y planificación física.

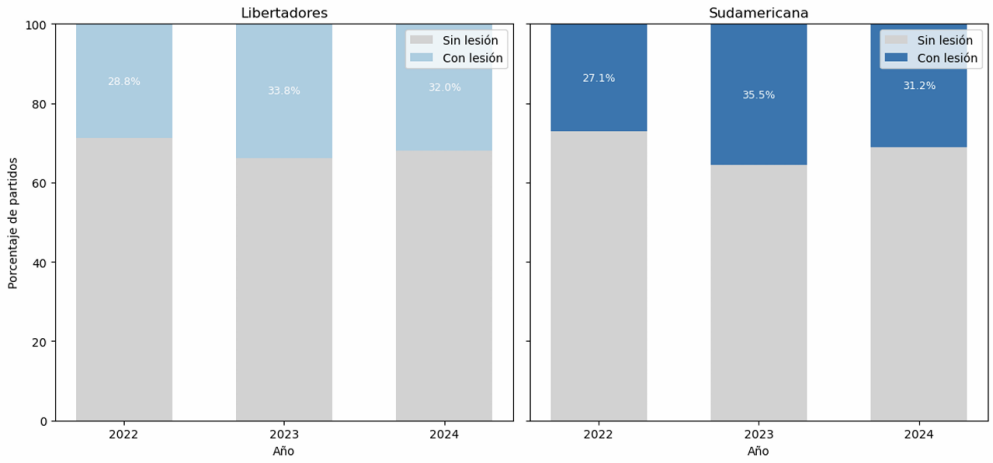

Figura 31: % relativo de partidos con lesión por competición

5.4.3 TASA DE LESIONES POR MINUTOS JUGADOS

Tras analizar los valores absolutos y relativos de lesiones por año y competición, se calcula ahora una métrica estandarizada que permite **comparar el riesgo de lesión ajustado por tiempo de juego**: la tasa por **1.000 minutos jugados**. Esta medida elimina el sesgo derivado del número variable de partidos y permite identificar tendencias reales en la incidencia de lesiones y los resultados se muestran en la Tabla 4.

Tipo	Valor absoluto	Cada x minutos que se produce	Lesiones / 1000 min
Total	363	225.1 min	4.4
Visitante	158	517.2 min	1.9
Local	205	398.6 min	2.5

Tabla 4: Resumen estadístico de lesiones

Antes hemos analizado el número absoluto de partidos con al menos una lesión. Ahora, al calcular lesiones por minuto jugado, obtenemos más información, representada en la Figura 32. Aunque **2023 sigue siendo el año con mayor incidencia**, ahora la Copa Libertadores

supera a la Sudamericana en tasa de lesiones por minuto. Esto tiene sentido ya que Libertadores es más competitiva, reúne equipos de élite y genera una mayor presión física y táctica, aumentando el estrés y riesgo de lesiones. Además, en 2024 Sudamericana presenta una tasa más alta que Libertadores, lo que lleva a examinar en detalle por fases.

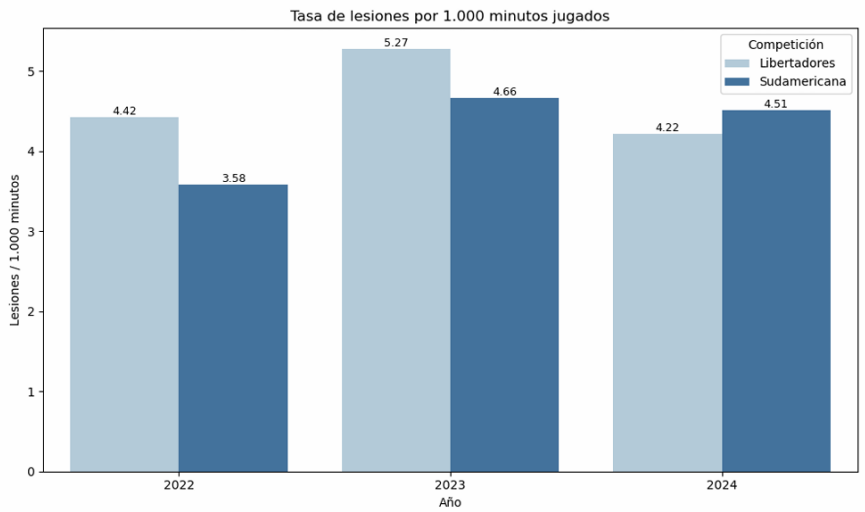

Figura 32: Tasa de lesiones por competición y año

Al desglosar por fases en la Figura 33, se confirma la hipótesis: la introducción de playoffs y calendario más comprimido aumenta significativamente las lesiones. En Sudamericana 2023, la **tasa en playoffs fue de 4.7 lesiones por 1.000 min**, y en 2024 de 22.2, aunque este último dato está sobredimensionado ya que como se mencionó en el apartado 5.3.2, solo hay registros de 3 de los 16 partidos por lo que no es representativo. No obstante, queda claro que los **playoffs incrementan el riesgo**, tanto por su exigencia física como por impacto en fases previas al comprimir entrenamientos y recuperación.

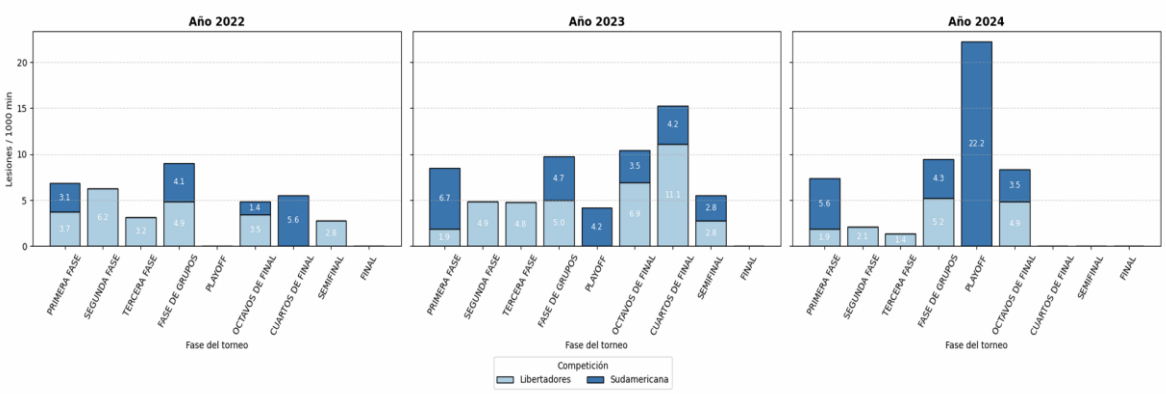

Figura 33: Tasa de lesiones por fase del torneo y competiciones

Tras detectar que la tasa de lesiones en la copa Sudamericana 2024 es especialmente alta, se analiza por condición de juego para identificar causas (Figura 34**)**. Se confirma que, **los jugadores locales tienen mayor tasa de lesión que los visitantes**, destacando que Sudamericana 2024 presenta una tasa de 3.22 para locales, frente a solo 1.45 para visitantes**,** la más baja del periodo. Por tanto, que Sudamericana tenga mayor tasa de lesión que Libertadores en 2024, se debe únicamente al incremento de lesiones locales.

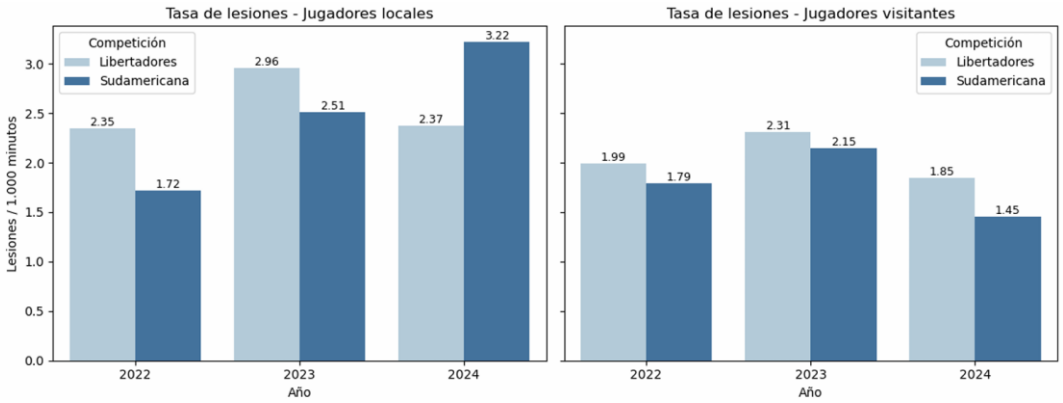

Figura 34: Tasa de lesiones por condición de jugador, año y competición

Se explora si las lesiones tienen relación directa con las victorias, es decir, a más victorias, más esfuerzo físico y por tanto mayor probabilidad de lesión. Los porcentajes de victorias en la Copa Sudamericana son 56%, 65% y 60% para locales y 40%, 29% y 30% para visitantes en las temporadas 2022, 2023 y 2024 respectivamente, lo que descarta la hipótesis

ya que 2024 debería haber tenido un porcentaje de victorias locales considerablemente más alto que años previos [50].

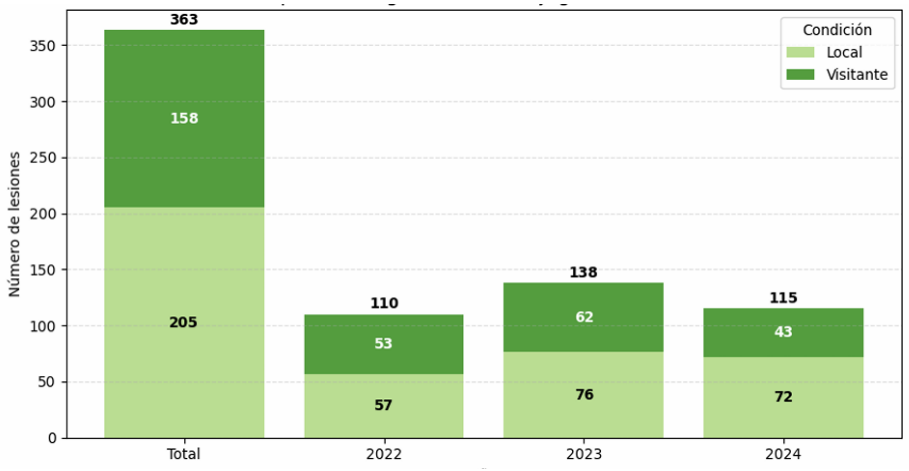

Figura 35: Lesiones local vs visitante por año

La Figura 35 confirma que los locales presentan siempre más lesiones, por lo que analizamos en profundidad para encontrar posibles causas. Viendo la proporción mensual de lesiones de locales vs visitantes en la figura 36, observamos que todos los meses a excepción de agosto y septiembre tienen mayor tasa local que visitante. Esto se puede deber a una menor frecuencia de partidos lo que distorsiona los datos al tener una baja representatividad o que, al tratarse de fases finales, ha condiciones que igualan el riesgo de lesión en ambos.

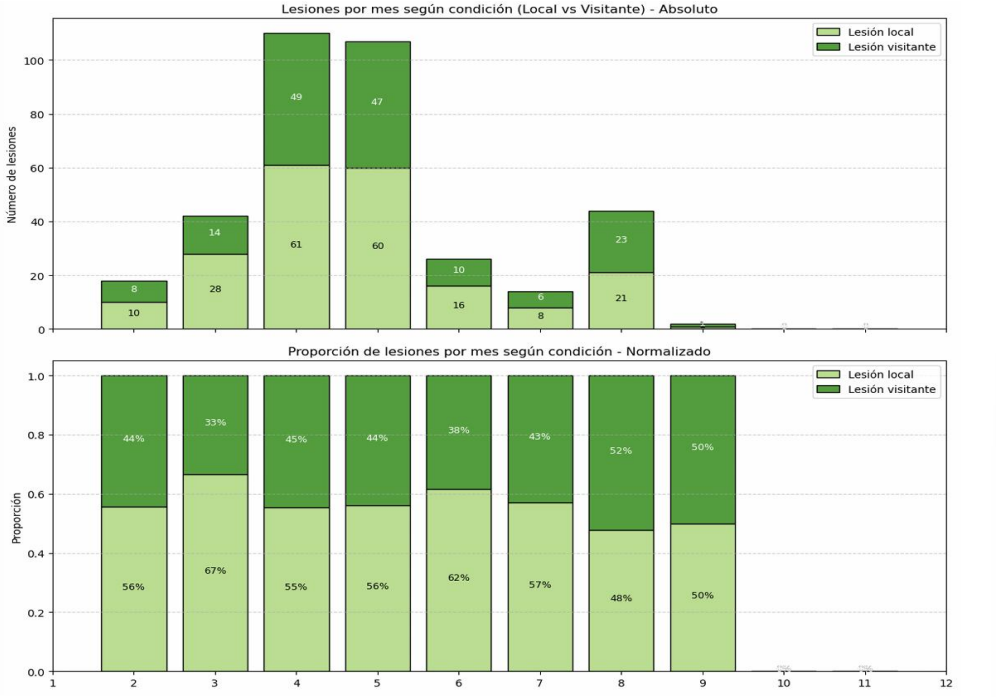

Figura 36: Lesiones absolutas y relativas de local vs visitante por mes

Para evaluar la representatividad de los datos, primero se analiza el número de partidos con lesión por mes (Figura 37). En septiembre solo hay 2 partidos con lesión (8%), por lo que se descarta ese dato, pero en agosto se registran 34 partidos (34%), en línea con otros meses, por lo que sí se considera relevante explorar por qué en fases finales las lesiones entre locales y visitantes tienden a igualarse.

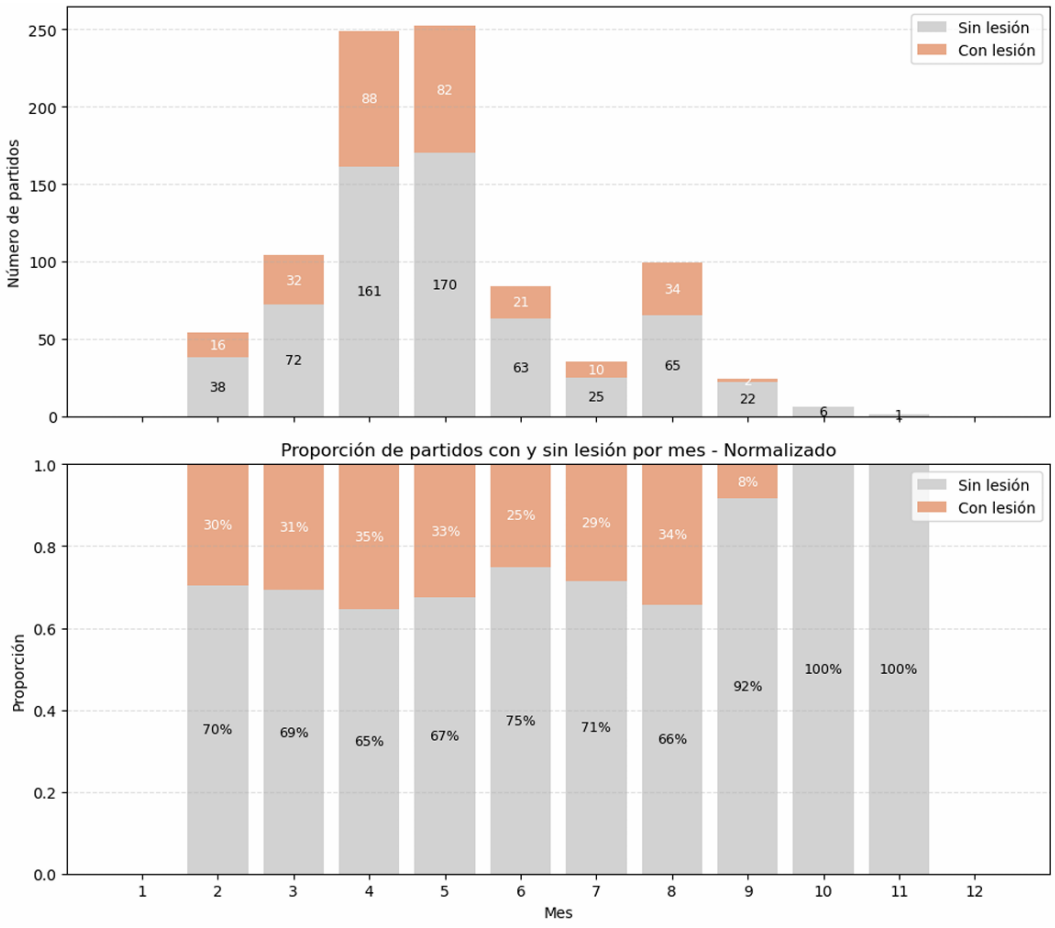

Figura 37: Número partidos por mes con lesión vs sin lesión

Una posible explicación al mayor número de lesiones en jugadores locales es la presión añadida de jugar en casa. Contar con el apoyo del público genera mayores expectativas y puede llevar a los jugadores a exigirse más físicamente, aumentando el riesgo de lesión. Además, los visitantes, presionados por el ambiente, tienden a jugar de forma más agresiva, lo que incrementa el número de faltas y contactos físicos.

Este planteamiento está respaldado por un estudio durante el periodo COVID-19, que comparó partidos con público frente a partidos sin espectadores. Los resultados muestran que cuando hay afición en las gradas, el equipo local domina más el juego (más remates) y los visitantes cometen más faltas y reciben más tarjetas. En ausencia de público, estos efectos desaparecen, lo que confirma que la afición influye tanto en el esfuerzo del equipo local como en la conducta más agresiva del rival [51].

Este patrón encaja con el análisis mensual (Figura 36). En los primeros meses del torneo, con mayor presencia de afición local, las lesiones se concentran en los equipos anfitriones. Sin embargo, en agosto y septiembre, ya en fases eliminatorias, las lesiones se reparten más entre locales y visitantes. Esto puede deberse a que en partidos importantes (como octavos, cuartos…) las aficiones están más equilibradas, lo que reduce el efecto de presión y sesgo local visto en las fases iniciales.

5.4.4 FACTORES GEOGRÁFICOS

En esta sección se estudian dos variables climáticas que podrían influir en el riesgo de lesión: **la altitud del estadio** y **la temperatura ambiente durante el partido**. En ambos casos, se observa en un primer análisis que existen ciertos rangos donde la tasa de lesiones por cada 1.000 minutos jugados es más alta. Por ejemplo, la altitud intermedia (1.001–2.500 m) muestra un pico de lesiones en 2023 (ver Figura 38), y las temperaturas entre 18 y 21 °C presentan una tasa considerablemente más elevada que el resto (ver Figura 39).

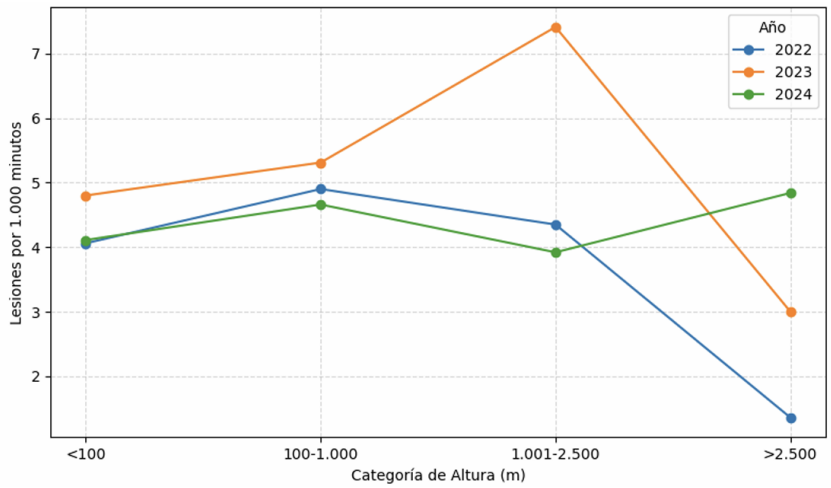

Figura 38: Tasa de lesiones / 1000 min según altura y año

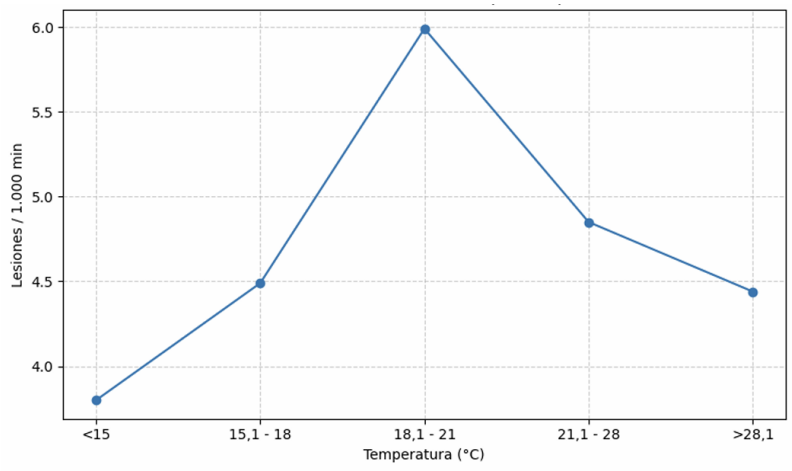

Figura 39: Tasa de lesiones por temperatura

Sin embargo, al analizar la **proporción de partidos con al menos una lesión**, los resultados se estabilizan entre todas las categorías de altitud y temperatura (ver Figura 40 y Figura 41). Es decir, **estos factores no parecen influir en la probabilidad de que ocurra una lesión en un partido,** aunque podrían estar relacionados con una mayor cantidad de lesiones cuando estas ocurren.

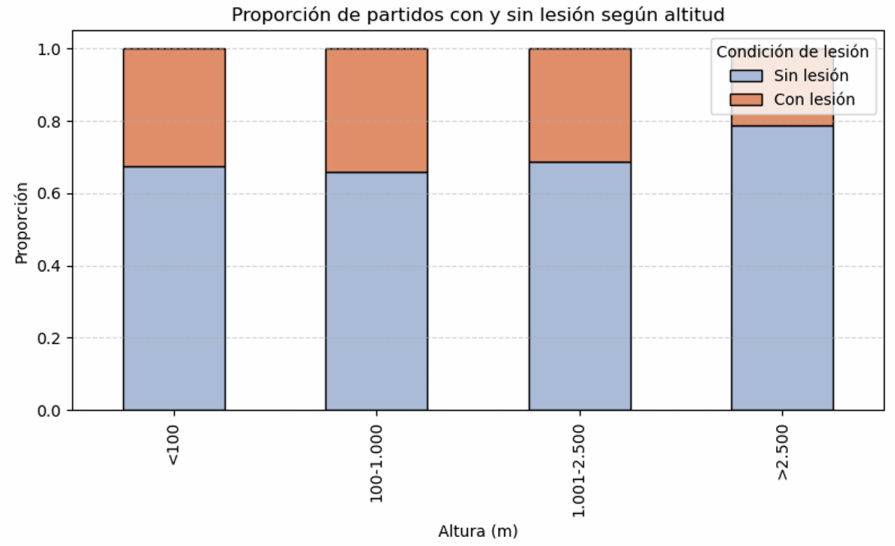

Figura 40: Proporción partidos con y sin lesión por altura

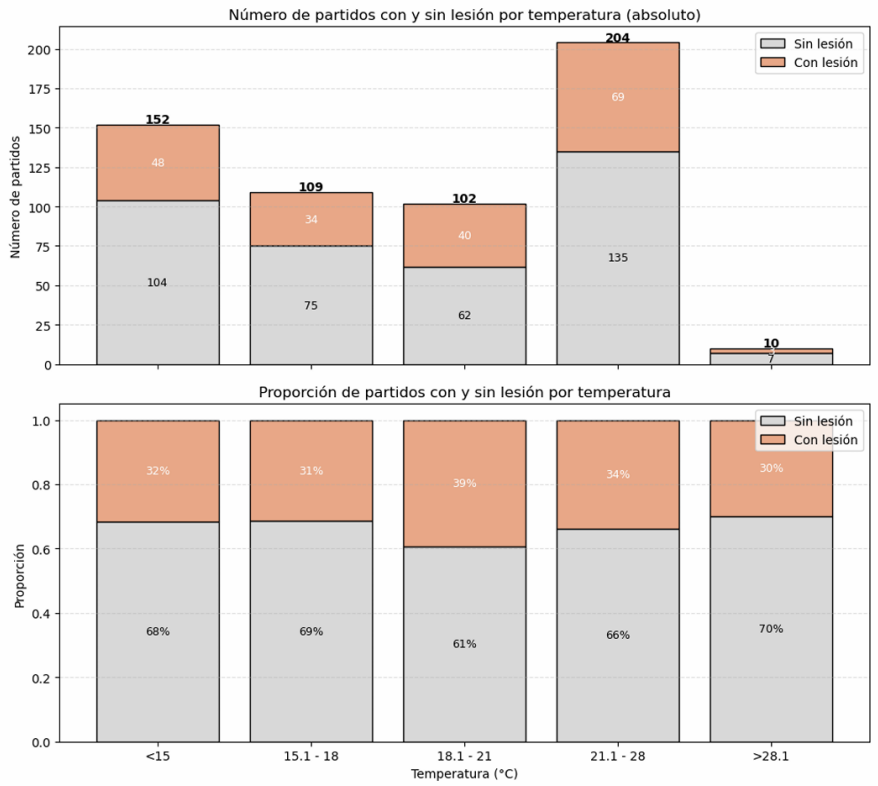

Figura 41: Valores absolutos y relativos de partidos con lesión por temperatura

Para confirmar si estas variables tienen influencia se utiliza la prueba de **chi-cuadrado**, que permite evaluar si los valores de una variable dependen o no de otra, con el objetivo de determinar si existe relación entre las dos variables [52]. Para ello se analiza si altura o temperatura influyen en la ocurrencia de lesiones donde la **hipótesis nula (H_0)** plantea que **no existe relación entre las categorías** climáticas y la presencia de lesiones. Se obtienen valores de p=0.0860 y p=0.7238 para altura y temperatura respectivamente, por lo que **no se rechaza la H_0**, confirmando la independencia de las variables.

Por tanto, ambas variables pueden tener algún impacto específico en contextos concretos (como estadios particulares o condiciones extremas), pero no parecen ser factores directamente relevantes a nivel general en este dataset.

Para cerrar el análisis de variables geográficas, se estudia la **distancia recorrida por el equipo visitante**. Aunque la proporción de partidos con lesión se mantiene estable entre

categorías (alrededor del 30 % en todas), al observar las tasas de lesiones por condición de juego y año aparecen patrones interesantes. En la Figura 42 se observa como en **2023 y 2024**, **las lesiones locales aumentan** cuando el **visitante ha recorrido más de 1000km,** mientras que las visitantes disminuyen.

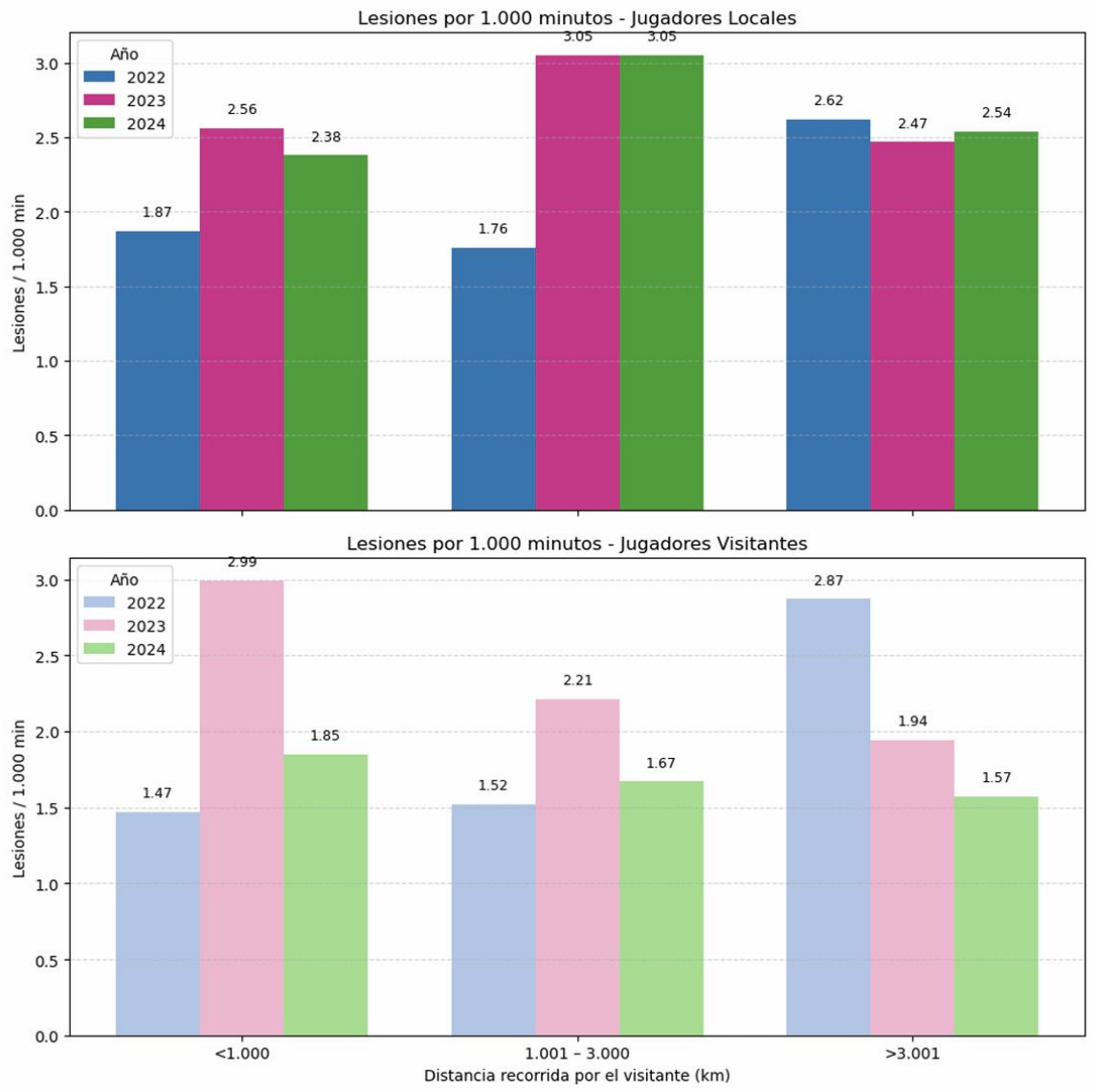

Figura 42: Tasa de lesiones según distancia recorrida por el equipo visitante

Este comportamiento no se repite en 2022, donde el aumento de distancia se asocia a una mayor tasa de lesiones tanto en locales como en visitantes, lo cual resulta más intuitivo (a mayor distancia, mayor cansancio general y, por tanto, más lesiones). El cambio de patrón

observado a partir de 2023 puede estar relacionado con el nuevo calendario competitivo (analizado el punto 5.4.2) que comprime más los partidos y reduce los tiempos de descanso. Este contexto, combinado con desplazamientos largos, podría haber amplificado el desgaste físico en los equipos.

Estos resultados refuerzan la hipótesis de que el **cansancio del visitante tras un viaje largo** puede derivar en un juego más impreciso o agresivo del visitante, lo que **incrementa el riesgo de lesión en los jugadores locales**.

Para cerrar el análisis geográfico, se estudia la **influencia de país anfitrión (**se ha estudiado la variable "ciudad", pero no se han encontrado conclusiones que aporten valor adicional). Como es de esperar, la Figura 43 muestra que Brasil y Argentina concentran el mayor número total de casos, tanto a nivel global como por años. Sin embargo, al observar la proporción de partidos con al menos una lesión (Figura 44), el orden cambia y países como Venezuela, Colombia o Ecuador lideran.

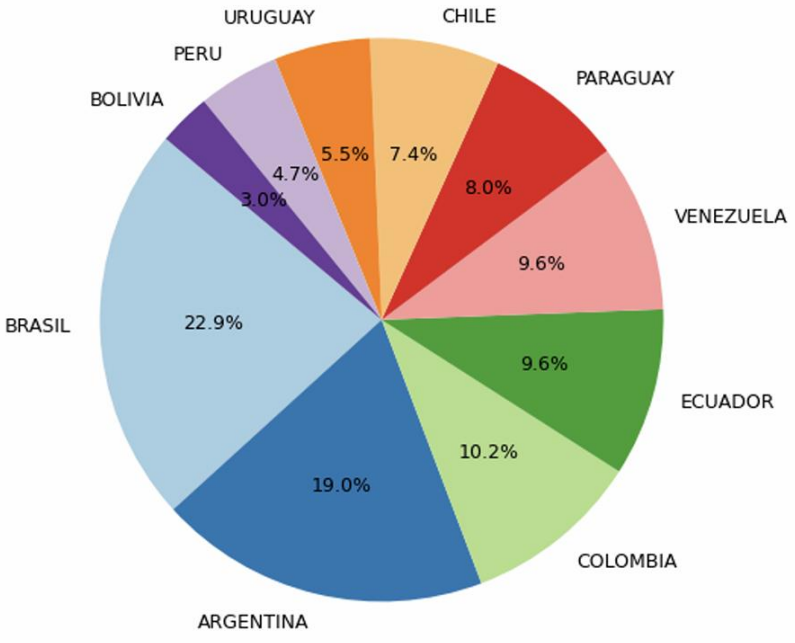

Figura 43: Distribución de lesiones totales por país

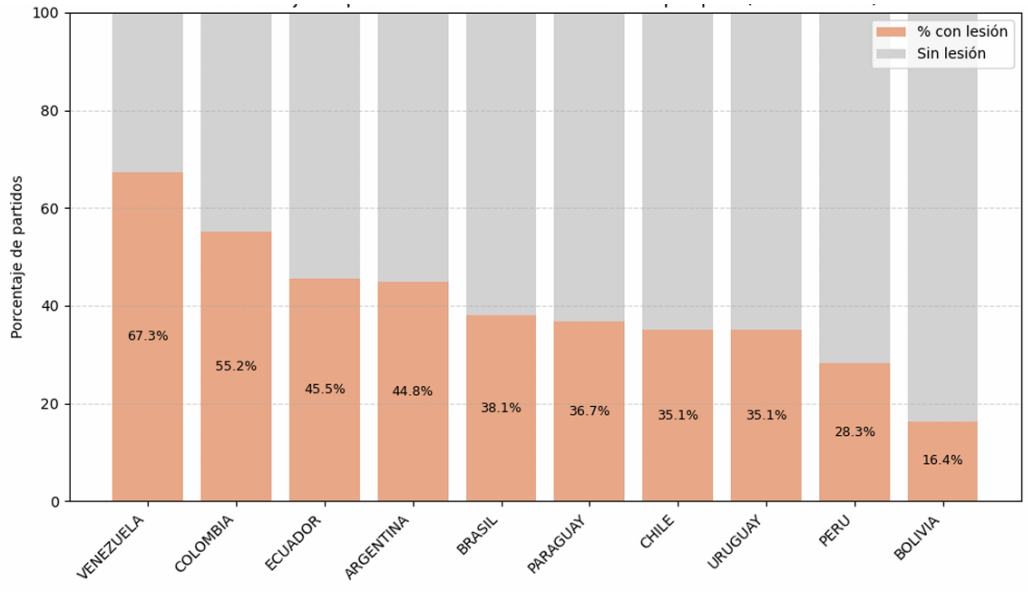

Figura 44: Porcentaje de partidos con y sin lesión por país

	Partidos totales			Partidos con lesión (%)		
	2022	**2023**	**2024**	**2022**	**2023**	**2024**
Brasil	74	76	68	36.5	43.4	33.8
Argentina	54	56	44	31.5	50.0	54.5
Colombia	20	28	19	60.0	50.0	57.9
Ecuador	30	27	20	36.7	40.7	65.0
Venezuela	20	15	17	85.0	80.0	35.3
Paraguay	25	25	29	32.0	40.0	37.9
Chile	25	24	28	28.0	37.5	39.3
Uruguay	16	18	23	56.2	38.9	17.4
Perú	22	22	16	18.2	40.9	25.0
Bolivia	25	18	24	0.0	27.8	25.0

Tabla 5: Número de lesiones por país y año

En el caso de **Venezuela**, las cifras altísimas se deben a la incidencia en **2022 (85 %) y 2023 (80 %),** aunque en 2024 se observa un fuerte descenso (Tabla 5). Este cambio puede estar relacionado con las inversiones realizadas en 2023 por la CONMEBOL en los estadios venezolanos para mejorar sus condiciones [53].

Colombia, por su parte, mantiene una incidencia alta y constante en los tres años analizados (Tabla 5), lo que podría estar relacionado con **el mal estado de sus estadios**. Según denuncias de jugadores profesionales, algunas instalaciones presentan condiciones inferiores a las de estadios de cuarta división española, algo difícil de justificar en el contexto del fútbol profesional [54].

Finalmente, otro aspecto relevante es la distancia recorrida por los visitantes ya que, por su ubicación, Colombia y Venezuela obligan a muchos rivales a desplazarse más km que para otros países.

.

Capítulo 6. MODELO PREDICTIVO DE LESIONES

6.1 INTRODUCCIÓN AL MACHINE LEARNING

6.1.1 ¿QUÉ ES EL MACHINE LEARNING? HISTORIA Y EVOLUCIÓN

El aprendizaje automático o machine learning (ML) es una rama de la inteligencia artificial que permite a los sistemas aprender automáticamente a partir de datos y mejorar su rendimiento sin necesidad de ser programados explícitamente. Su origen se remonta a mediados del siglo XX, con hitos como el Perceptrón de Rosenblatt (1958) y los trabajos de Arthur Samuel sobre aprendizaje de máquinas en juegos. Sin embargo, ha sido en las últimas décadas, con la aparición de grandes volúmenes de datos y capacidad computacional cuando el ML ha adquirido un papel protagonista [55].

Existen dos grandes enfoques dentro del aprendizaje automático: el **aprendizaje supervisado y el no supervisado.** El primero consiste en entrenar modelos a partir de datos etiquetados, donde la variable objetivo es conocida (como en problemas de clasificación o regresión). El segundo, en cambio, busca encontrar patrones o estructuras ocultas en los datos sin una variable objetivo-específica [56].

6.1.2 CASOS DE USO DEL ML EN SALUD, DEPORTE Y PREDICCIÓN DE EVENTOS

El ML ha demostrado un enorme potencial para transformar la medicina y el rendimiento deportivo. En el ámbito clínico, se ha aplicado con éxito al diagnóstico de enfermedades mediante imágenes médicas, a la predicción de riesgo cardiovascular, a la detección temprana de cáncer, o a la monitorización de constantes vitales mediante wearables [57].

Más allá, también ha sido clave para la optimización de tratamientos personalizados en oncología. Por ejemplo, se han desarrollado modelos que predicen la respuesta de los pacientes a determinadas terapias basados en perfiles genómicos y biomarcadores.

Asimismo, su uso en el análisis de imágenes radiológicas ha permitido detectar con mayor precisión lesiones o tumores en etapas tempranas, mejorando el pronóstico clínico [58]

En el contexto deportivo, el ML ha permitido avances importantes como la individualización de cargas de entrenamiento, la predicción de rendimiento, la evaluación de fatiga o la prevención de lesiones musculares a partir de patrones históricos [59].

6.2 TIPOS DE MODELOS DE CLASIFICACIÓN DE ML

6.2.1 PREDICCIÓN DE VARIABLES CATEGÓRICAS: CLASIFICACIÓN

Los modelos de clasificación permiten asignar a una observación una categoría entre varias posibles. En el caso de este trabajo, se trata de una **clasificación binaria**: predecir si se producirá (1) o no (0) al menos una lesión durante un partido. Existen distintos tipos de clasificadores [60].

Entre los modelos más comunes destacan: Regresión logística, Random Forest, Extreme Gradient Boosting (XGBoost), Support Vector Machine (SVM) y K-Nearest Neighbors (KNN) [60].

En este proyecto, se opta por usar **Regresión Logística**, **Random Forest y XGBoost** por su buen desempeño en entornos con datos desbalanceados y relaciones no lineales, características del dataset. Además, un análisis reciente en predicción de lesiones deportivas refuerza esta decisión, identificando los modelos de **Random forest y XGBoost** como los **más usados con un 36% y 19%,** frente a un uso de menos del 10% para el resto de los modelos en este tipo de aplicaciones [59].

6.2.1.1 Regresión logística

La regresión logística es un modelo estadístico ampliamente utilizado para problemas de clasificación binaria. A diferencia de la regresión lineal, **su objetivo no es predecir un valor continuo, sino la probabilidad de pertenencia a una clase.** En este caso, se utiliza para estimar la probabilidad de que un partido de fútbol con determinadas características derive

en al menos una lesión. Es un modelo interpretable, lo que permite analizar la influencia relativa de cada variable predictora sobre el resultado.

La regresión logística transforma una combinación lineal de variables independientes a una probabilidad mediante la función sigmoide y se expresa con la Ecuación 1.

$$P\,(Y = 1) = \frac{1}{1 + e^{-(\beta 0 + \beta 1 + \cdots + \beta nxn)}}$$

Ecuación 1: función sigmoide de regresión logística

donde P(Y=1) es la probabilidad de que ocurra una lesión, x_i son las variables predictoras, y βi los coeficientes ajustados durante el entrenamiento. La salida continua (entre 0 y 1) se transforma en una predicción binaria según un umbral de decisión [61].

La regresión logística funciona bien cuando existe una relación aproximadamente lineal entre las variables predictoras y el log-odds (Ecuación 2) de la variable objetivo. Es especialmente útil cuando se desea interpretabilidad y cuando se cuenta con datasets de tamaño moderado y sin gran cantidad de ruido o multicolinealidad.

$$log - odds = \log(\frac{p}{1-p})$$

Ecuación 2: Logaritmo de la razón de probabilidades

6.2.1.2 Árboles de decisión y Random Forest:

El modelo de Random Forest es un algoritmo de ensamblado que **construye múltiples árboles de decisión sobre subconjuntos aleatorios** del dataset y combina sus predicciones para mejorar la precisión y evitar el sobreajuste. Su principio básico consiste en generar una "votación" entre árboles independientes, reduciendo la varianza del modelo individual [56]. Cada árbol se entrena sobre una muestra diferente de datos (mediante bagging) y en cada nodo se consideran aleatoriamente solo un subconjunto de características, lo cual fortalece su capacidad para generalizar.

Matemáticamente, la predicción final se obtiene por mayoría de votos (clasificación) o promedio (regresión) de las predicciones individuales de los árboles. Esta técnica resulta especialmente útil en **contextos con muchas variables y relaciones no lineales**, donde existen interacciones complejas entre factores geográficos, temporales y contextuales y además donde hay variables con valores nulos.

6.2.1.3 XGBoost (Extreme Gradient Boosting):

El modelo XGBoost es un algoritmo de ensamblado basado en la técnica de *boosting*, que consiste en **construir modelos de forma secuencial donde cada nuevo árbol corrige los errores del anterior.** A diferencia de Random Forest, que crea árboles en paralelo, XGBoost optimiza la función objetivo minimizando una pérdida mediante gradiente descendente, lo que permite una mejora iterativa del rendimiento del modelo.

Este modelo destaca por su alta capacidad predictiva, velocidad de entrenamiento y flexibilidad en el **manejo de datos incompletos**, **variables categóricas** y **relaciones no lineales**. Tiene alta capacidad para priorizar variables relevantes automáticamente y manejar de manera eficiente grandes volúmenes de datos [62].

6.2.2 MÉTRICAS DE EVALUACIÓN

Se van a utilizar principalmente 5 métricas para la evaluación de los modelos, pero antes de explicarlas se debe tener en cuenta las siguientes siglas para poder entender las fórmulas matemáticas de cada métrica:

- TP = True Positive or Verdadero Positivo
- FP = False Positive o Falso Positivo
- TN = True Negative o Verdadero Negativo
- FN = False Negative o Falso Negativo

- **Accuracy**: Mide el **porcentaje de predicciones correctas realizadas por el modelo sobre el total de casos** y su cálculo se muestra en la Ecuación 3. En problemas desbalanceados, como es este caso ya que la mayoría de los partidos no presentan

lesiones, esta métrica puede resultar engañosa: un modelo que predice siempre "no lesión" puede tener una accuracy alta, pero no estaría detectando los casos interesantes que son los partidos con lesión [63].

$$Accuracy = \frac{TP + TN}{Total}$$

Ecuación 3: Fórmula de Accuracy

- **Precision:** Evalúa la **calidad de las predicciones positivas**, midiendo qué proporción de los partidos que el modelo predice como "con lesión" realmente presentan una lesión, Ecuación 4. Una alta precisión implicará pocas falsas alarmas, lo cual es deseable, pero no suficiente por si se sacrifican muchos verdaderos positivos, por ello, se dará más importancia al recall [63].

$$Precision = \frac{TP}{TP + FP}$$

Ecuación 4: Fórmula de precision

- **Recall o sensibilidad**: Representa la **capacidad del modelo para identificar correctamente los casos positivos,** Ecuación 5. En este trabajo será la métrica prioritaria, ya que el objetivo es detectar el mayor número posible de partidos con al menos una lesión. Esto se justifica por la importancia clínica y preventiva de anticipar estos casos, aunque se den más falsos positivos [63].

$$Recall = \frac{TP}{Tp + FN}$$

Ecuación 5: Fórmula recall

- **F1-score**: Media armónica entre precisión y recall (Ecuación 6) y es útil cuando se necesita un equilibrio entre ambas, especialmente en escenarios con clases desbalanceadas. Al penalizar los extremos (por ejemplo, alta precisión, pero bajo recall), el F1-score proporciona una evaluación más robusta del modelo [63].

$$F1 - score = \frac{Precision * Recall}{Precision + Recall}$$

Ecuación 6: Fórmula F1-score

- **AUC-ROC (área bajo la curva ROC)**: Capacidad del modelo para distinguir entre clases positivas y negativas en todos los umbrales posibles. La curva ROC traza la relación entre la tasa de verdaderos positivos (Ecuación 7) y la tasa de falsos positivos (Ecuación 8) para distintos umbrales. El área bajo esta curva (AUC) proporciona un valor resumen de la discriminación global del modelo [63].

$$TPR\ (o\ recall) = \frac{TP}{TP + FN}$$

Ecuación 7: Fórmula True Positive Rate o Tasa de Verdaderos Positivos

$$FPR\ = \frac{FP}{FP + TN}$$

Ecuación 8: Fórmula False Positive Rate o Tasa de Falsos Positivos

6.3 *MODELO DE PREDICCIÓN GLOBAL DE LA VARIABLE LESIÓN*

6.3.1 PREPARACIÓN DE DATOS Y BASES UTILIZADAS

El modelo predictivo desarrollado tiene como objetivo estimar la probabilidad de que se produzca al menos una lesión durante un partido. Para ello, se define una variable binaria denominada LESION, construida a partir de la columna que representa el número total de lesiones ocurridas en cada partido, asignando un 1 cuando se ha dado al menos una lesión y 0 en caso contrario, permitiendo reformular el problema como una clasificación binaria.

Durante el proceso de preparación del dataset para el entrenamiento de los modelos, se han aplicado diferentes criterios para la eliminación de columnas. Se han eliminado aquellas que presentan un número elevado de valores nulos (más del 40 %), como es el caso de la variable IDA_VUELTA. También se excluyen variables con más de 10 categorías, como las de los

nombres de equipos y las ciudades del partido ya que su conversión a variables dummies introduce demasiada complejidad. Finalmente, se descartan columnas que tienen relación directa con la variable objetivo, como son las variables que indican el diagnóstico, la localización y el minuto de la lesión ya que están directamente relacionadas con la presencia de una lesión.

A partir de estos criterios, se generan dos versiones distintas del dataset:

- **Base con nulos**: conserva la variable temperatura (con un ~30% de valores nulos) y se utilizará para entrenar los modelos que toleran internamente los nulos, como Random Forest y XGBoost.
- **Base sin nulos**: elimina las filas de partidos con valores nulos, perdiendo ~30% de los datos y se usará para los modelos que requieren inputs sin nulos, como la regresión logística y stacking.

Además, el dataset presenta un **desbalance de clases significativo** lo que representa un desafío en la tarea de clasificación, por lo que se priorizarán métricas como el recall, la F1-score o el área bajo la curva ROC (AUC), explicadas el punto 6.2.2, a la hora de evaluar los modelos.

6.3.2 MODELO BASE

Para entrenar el modelo base, se ha utilizado la versión del dataset sin valores nulos para regresión logística y con nulos para XGBoosting y Random Forest, y se han establecido los parámetros por defecto de los modelos.

6.3.2.1 Mejoras aplicadas:

Se han implementado diferentes técnicas de optimización y mejora de los modelos, tanto en el preprocesamiento como en el ajuste interno de cada algoritmo. Los resultados más relevantes se muestran más adelante en los resultados, en la Tabla 6 del apartado punto 7.1.

Como se mencionó en el apartado 6.3.1, para **regresión logística** usamos una base de datos incluyendo la variable temperatura a cambio de la pérdida del 33% de los datos por valores nulos. Probamos quitando dicha variable, ganando de vuelta estos datos, pero los resultados fueron peores por lo que nos mantenemos con la hipótesis inicial de priorizar más variables predictoras por encima de más datos [18].

Además, al tener un desbalance de clases elevados, se implementan técnicas como **class_weight = balanced**, que ajusta automáticamente los pesos de las clases para compensar o **scale_pos_weight**, que compensa asignando mayor peso a la clase positiva para que el desajuste no afecte gravemente al modelo.

Una vez definido el conjunto de predictores, se ha aplicado **regularización** para evitar el sobreajuste. En el caso de la regresión logística, se han evaluado tres técnicas distintas: **L1 (Lasso)**, que tiende a eliminar variables no relevantes; **L2 (Ridge)**, que penaliza grandes coeficientes; y **ElasticNet**, que combina ambas penalizaciones. Estas estrategias han sido implementadas mediante el parámetro penalty de la clase LogisticRegression de scikit-learn, y han sido fundamentales para mejorar la generalización del modelo ante datos ruidosos o multicolineales. Tras probar diferentes valores, obtenemos los mejores resultados con ElasticNet, definida por los parámetros c y l1_ratio [18].

Para el modelo de XGBoost, se comienza con el XGBClassifier como modelo base con el dataset sin nulos. Posteriormente se prueba con el modelo Pipeline, que permite integrar de manera ordenada todas las etapas previas al entrenamiento, como la selección de variables, escalado, imputación de valores nulos o transformación de variables categóricas. Tras una mejora significativa, se aplica directamente esta técnica para el modelo random forest.

Adicionalmente, se ha llevado a cabo un proceso sistemático de ajuste de hiper parámetros utilizando **GridSearchCV** y **RandomizedSearchCV**, técnica que ha permitido evaluar más de 250 combinaciones diferentes de parámetros e hiper parámetros para identificar la que optimiza el rendimiento del modelo. Estos parámetros varían en función de los modelos y son los siguientes:

- **Regresión logística**: En el modelo de regresión logística con regularización ElasticNet, se ha ajustado el parámetro C, que representa el inverso de la fuerza de regularización. Un valor alto de **C** implica una menor penalización y mayor flexibilidad del modelo, mientras que un valor bajo fuerza a los coeficientes a ser más pequeños, lo que puede reducir el sobreajuste. Tras evaluar múltiples combinaciones, se ha encontrado que valores intermedios de C permiten un buen equilibrio entre varianza y sesgo. Además, se ha ajustado el parámetro **l1_ratio**, que determina la proporción entre regularización L1 (Lasso) y L2 (Ridge). Valores **cercanos a 0 favorecen Ridge** (más estable con muchas variables poco informativas), mientras que **valores altos favorecen Lasso** (más útil cuando se quiere seleccionar automáticamente variables relevantes) [18].

- **Random Forest:** El primero es **n_estimators**, que indica el número de árboles en el bosque. Aumentar este valor generalmente mejora la estabilidad del modelo, aunque con mayor coste computacional. Se ha optado por valores en el rango de 100 a 300. También se ha ajustado **max_depth**, que controla la profundidad máxima de cada árbol. Limitar esta profundidad reduce el riesgo de sobreajuste, especialmente útil con datasets ruidosos o con pocas observaciones. Otro parámetro importante ha sido **min_samples_split**, que determina el número mínimo de muestras necesarias para dividir un nodo; valores más altos hacen que los árboles sean más conservadores, evitando particiones sobre ajustadas [18].

- **XGBoost:** El primero es **n_estimators**, al igual que en Random Forest, cuyo aumento tiende a mejorar el rendimiento hasta cierto punto, aunque con riesgo de sobreajuste si no se controla la regularización. También se ha ajustado **max_depth**, que limita la profundidad de los árboles base y ayuda a prevenir modelos demasiado complejos. El parámetro **learning_rate** ha sido especialmente relevante: valores bajos (como 0.05 o 0.1) permiten aprender de forma más gradual, requiriendo más árboles, pero favoreciendo una mejor generalización. Se ha afinado también **subsample**, que define la proporción de datos que se usan en cada iteración de

boosting; valores menores a 1.0 introducen aleatoriedad, lo que puede mejorar la robustez del modelo [18].

Buscando maximizar el recall, se ha probado realizar el **ajuste del umbral de decisión** para todos los clasificadores. Por defecto, se clasifica como lesión cualquier caso cuya probabilidad supere el 50 % por lo que se evalúan umbrales alternativos. Esta técnica se puede observar en la Figura 45, para el caso de regresión logística, en el que podemos observar cómo efectivamente umbrales más bajos mejoran el recall considerablemente a costa de perjudicar el accuracy. En este caso elegimos el umbral 0.5 ya que no podemos permitir tener un accuracy < 0.5 (implica tener un modelo inferior al azar).

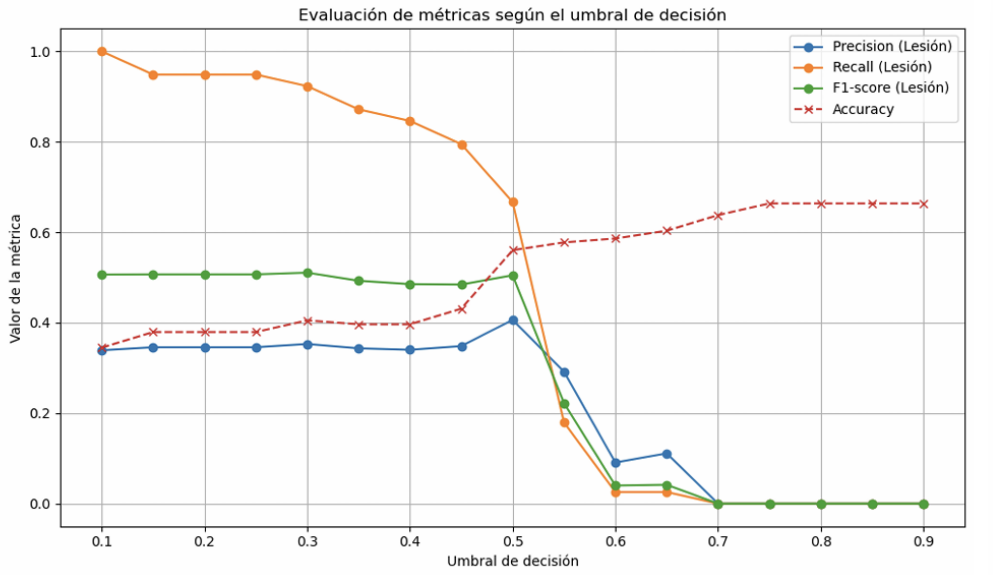

Figura 45: Métricas según umbral en regresión logística

Finalmente, a la hora de evaluar si estas técnicas mejoraban o no los modelos originales y los que evolucionaban con el proceso, nos hemos centrado en **aumentar F1-score y recall, sin perjudicar un accuracy por debajo de 0.55** (la del modelo base).

6.3.3 MODELOS DE COMBINACIÓN (ENSEMBLE)

Una vez definidos y optimizados los tres modelos base (Regresión Logística, Random Forest y XGBoost), se ha evaluado la posibilidad de mejorar el rendimiento general mediante su

combinación. Para ello, se implementan dos estrategias ampliamente utilizadas en machine learning: **Voting Ensemble** y **Stacking**, cuyos resultados se muestran en la Tabla 7 en el apartado 7.1.

6.3.3.1 Voting Ensemble

El método de Voting Ensemble consiste en combinar las predicciones de múltiples modelos mediante una votación ponderada. En este caso, se han empleado los tres modelos finales ajustados individualmente: Regresión Logística, Random Forest y XGBoost. La predicción final se obtiene como una media ponderada de las probabilidades individuales de cada modelo.

Para determinar la combinación óptima de pesos asignados a cada modelo, se ha aplicado un proceso de grid search sobre múltiples configuraciones. El objetivo ha sido maximizar el recall sin comprometer significativamente el resto de las métricas. El ensemble ponderado no demuestra una mejora global en el rendimiento sobre los modelos individuales.

6.3.3.2 Stacking

El stacking es un modelo que ha demostrado ser especialmente efectivo para mejorar la capacidad del sistema en contextos de datos ruidosos o desbalanceados, al permitir que el modelo meta aprenda patrones de error entre los clasificadores base y los corrija de forma automática. Es una técnica más sofisticada que **combina modelos mediante una arquitectura en dos niveles**. En la primera capa se utilizan los modelos base ya entrenados (Regresión Logística, Random Forest y XGBoost), que generan predicciones sobre los datos. Estas predicciones se convierten en nuevas variables de entrada para un **modelo meta**, que se encarga de aprender a partir de esas salidas cómo tomar la decisión final, esta estructura se puede observar en la Figura 46.

Se han probado tres algoritmos como modelo meta y los resultados se encuentran en la Tabla 7 en el apartado 7.1.

- Regresión logística: por su interpretabilidad y bajo riesgo de sobreajuste.

- Random Forest: por su capacidad de captura de interacciones complejas.
- XGBoost: por su alta precisión y capacidad de generalización.

Cada arquitectura se ha evaluado con **cross-validation y grid search,** repitiendo el proceso de ajuste de hiper parámetros como en los modelos individuales, buscando maximizar recall, F1 y AUC sin perjudicar accuracy. El modelo final obtenido es con regresión logística y su arquitectura interna se muestra en la Figura 47.

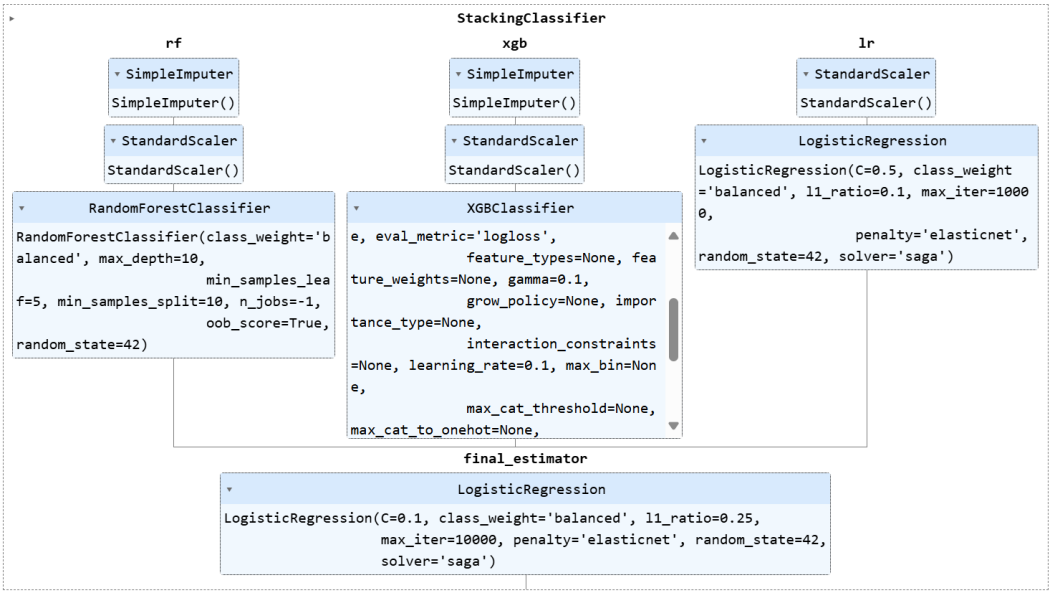

Figura 46: Estructura modelo stacking final

6.4 CLASIFICACIÓN POR TIPO DE LESIÓN (DIAGNÓSTICO / LOCALIZACIÓN)

Tras entrenar el modelo general y llegar a los 4 modelos finales, se explora una segunda vía: predicción del tipo de lesión, tanto desde el punto de vista del diagnóstico médico como de su localización anatómica. Este enfoque segmentado permite obtener información más precisa sobre las características de las lesiones potenciales, lo que resulta valioso para diseñar medidas preventivas específicas.

Debido a la alta cardinalidad de las variables diagnóstico y localización se han agrupado previamente en 13 tipos de diagnóstico y 5 localizaciones. Además, por la poca disponibilidad de datos también se aplica un filtro de frecuencia mínima, solo considerando para el análisis aquellos tipos de lesiones con más de 20 registros. Decisión que se toma para evitar problemas de sobreajuste y convergencia en los modelos.

También se impone una restricción cruzada en la elección de variables: cuando se está prediciendo el diagnostico, no se permite utilizar localización como variable predictora, y viceversa, ya que cada lesión tiene una localización y diagnostico correspondiente por lo que se relacionan linealmente. Pero, si se ha explorado la posibilidad de incluir el resto de las localizaciones o diagnósticos a la hora de su predicción. Es decir, si quiero predecir si habrá o no lesiones musculares, también incluyo los datos de si ha habido esguinces, contusiones… ya que, que haya habido ya una lesión en el partido puede indicar mayor probabilidad de que haya otra, o bien puede haber tipos de lesiones que se den juntas.

6.4.1 SELECCIÓN MANUAL DE VARIABLES

Se han aplicado técnicas de selección de características (feature selection) basadas en los métodos de cada modelo para identificar qué variables tienen mayor impacto predictivo. En el caso de los modelos basados en árboles como Random Forest y XGBoost, se ha utilizado el atributo feature_importances_ para evaluar la importancia relativa de cada variable, mientras que en el caso de la regresión logística se han analizado los coeficientes resultantes.

Los resultados de este análisis, en la Tabla 8, permiten entender mejor cada modelo, ajustando las variables que se utilizan para entrenarlos con el objetivo de reducir el ruido, mejorar la capacidad generalizadora y optimizar los resultados. Esta estrategia de selección diferencial por modelo también se traslada a la interfaz final del sistema, permitiendo al usuario seleccionar manualmente las variables contextuales más adecuadas según el modelo seleccionado [64].

Capítulo 7. ANÁLISIS DE RESULTADOS

7.1 MODELOS FINALES

Una vez entrenados y optimizados los tres modelos base; Regresión Logística, Random Forest y XGBoost. En la Tabla 6 se muestran los resultados más relevantes obtenidos para cada tipo de modelo a lo largo del proceso iterativo de ajuste y validación. Para cada uno se ha buscado potenciar una métrica concreta: **recall** en el caso de la Regresión Logística, **accuracy** en el XGBoost y **AUC** en Random Forest. Este enfoque ha permitido cubrir distintos puntos fuertes, clave para su posterior combinación.

Tipo de modelo	Accuracy	Precision	Recall	AUC	F1 - score
Reg. Logística inicial	0.6034	0.1111	0.0256	0.5354	0.0416
Reg. Logística con balance de clases	0.5517	0.3859	0.5641	0.5334	0.4583
Reg. Logística (sin temperatura)	0.5384	0.3373	0.4912	0.5364	0.4000
Reg. Logística L1 (Lasso)	0.5258	0.3620	0.5384	0.5194	0.4329
Reg. Logística L2 (Ridge)	0.5344	0.3584	0.4871	0.5214	0.4130
Reg. Logística L1+L2 (c=0.05, l1_ratio = 0,1)	**0.5775**	**0.4193**	**0.6666**	**0.5338**	**0.51485**
Reg. Log L1+L2 (sin temp, año y trimestre, coef = 0)	0.5275	0.3370	0.5263	0.5534	0.4109
XGBoost inicial	0.5775	0.3333	0.2564	0.4965	0.2898
XGBoost (randomized search y mejor hiperparametros)	0.5689	0.3513	0.3333	0.4385	0.4385
XGBoost (imputación de nulos)	0.6428	0.4166	0.3508	0.5020	0.3809

XGBoost (grid search y major hiperparametros)	0.6428	0.4200	0.3684	0.5489	0.3925
XGBoost (con Pipeline + grid search)	**0.6373**	**0.4151**	**0.3859**	**0.5467**	**0.4000**
Random Forest (pipeline + grid search inicial)	0.5495	0.3684	0.614	0.5642	0.4605
Random Forest (major combinacion hiperparam.)	0.6044	0.4026	0.5439	0.5544	0.4627
Random forest (u = 0.48)	**0.6044**	**0.4026**	**0.5439**	**0.5761**	**0.4627**

Tabla 6: Métricas modelos evaluados

En la Tabla 7 se presentan los resultados de los modelos combinados. El modelo de ensemble ponderado se ha construido asignando pesos [0.5, 2, 1] a los modelos base. Por otro lado, el modelo de stacking integra las predicciones individuales de los tres modelos base (RF, XGB, LR) como entradas para un nuevo clasificador final y se ha elegido regresión logística como regresor final.

Tipo de modelo	Accuracy	Precision	Recall	AUC	F1 - score
Reg. Logística final	**0.5775**	**0.4193**	**0.6666**	**0.5338**	**0.51485**
XGBoost final	**0.6373**	**0.4151**	**0.3859**	**0.5467**	**0.4000**
Random forest final	**0.6044**	**0.4026**	**0.5439**	**0.5761**	**0.4627**
Ensemble, w = [0,5;2,1]	0.5690	0.3778	0.4359	0.4965	0.4048
Stacking con XGB	0.5000	0.3699	0.6923	0.4680	0.4821
Stacking con Random Forest	0.4310	0.3412	0.7436	0.4902	0.4677
Stacking con Reg. Logística	**0.5862**	**0.4348**	**0.7692**	**0.6254**	**0.5556**

Tabla 7: Métricas modelos ensemble + stacking

El **modelo stacking final** (Figura 46) obtiene un **recall del 76.92 %**, el valor más alto alcanzado en todo el proyecto, y mejora también el **F1-score (0.5556)** y el **AUC (0.6254)**, a costa de una ligera caída en la accuracy, que se mantiene en niveles aceptables. Este resultado refleja una mejora global significativa frente a los modelos individuales, cumpliendo con el objetivo principal de maximizar la detección de lesiones reales.

Finalmente, la curva ROC comparativa (Figura 47) confirma visualmente estas mejoras, donde la línea negra correspondiente al modelo de stacking y se sitúa consistentemente por encima del resto, indicando una mejora en la capacidad de discriminación entre clases.

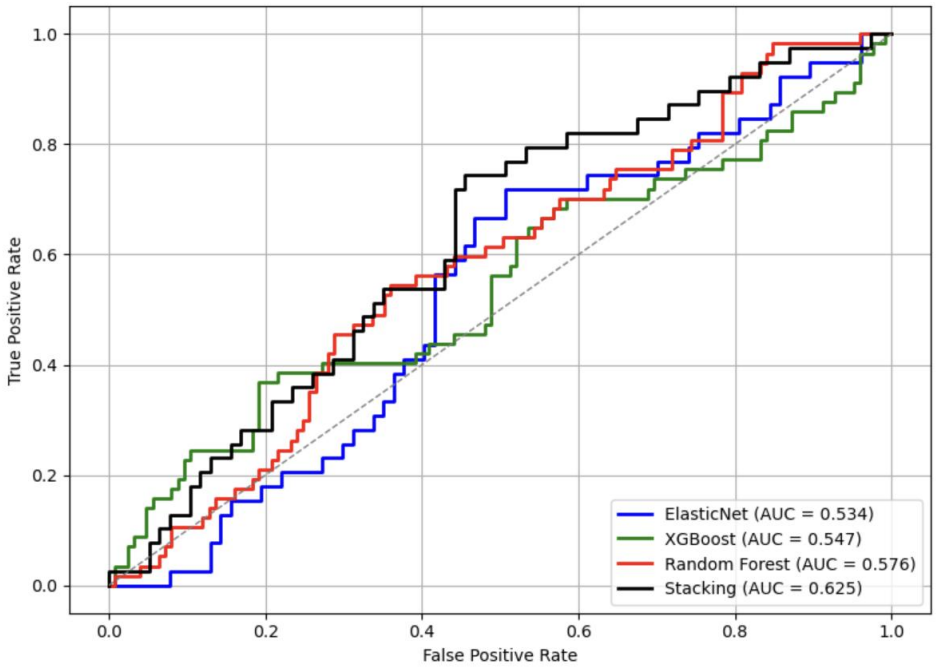

Figura 47: Curva ROC de los modelos finales

7.2 *VARIABLES MÁS RELEVANTES*

Los resultados de los análisis de variables relevantes mencionados anteriormente se encuentran en la Tabla 8.

Variable	Regresión-logística	XGBoost	Random forest
Team1 Bolivia	- 0.6164	0.1125	0.0614
Team1 Perú	- 0.0200	0.0464	0.0105
Team1 Brasil	- 0.1253	0.0211	0.0118
Team2 Bolivia	•	0.0235	0.0110
Team2 Uruguay	- 0.1083	0.0222	0.0077
Pais Bolivia	-	0.0977	0.0421
Pais Peru	- 0.1235	0.0283	0.0129
Pais Uruguay	- 0.2383	0.0161	0.0073
Pais Paraguay	- 0.1099	0.0247	0.0072
Pais Ecuador	0.1647	0.0283	0.0078
Distancia visitante	- 0.1587	0.0207	0.1434
Altura	- 0.0172	0.0210	0.1145
Temperatura	-	0.0211	0.0826
Mes	- 0.0748	0.0193	0.0748
Año	-	0.0200	0.035
Competición Sudamericana	- 0.0421	0.0173	0.0239
Fase de clasificación	- 0.2791	0.0224	0.0084
Fase octavos de final	- 0.2333	0.0356	0.0085

Tabla 8: Variables más relevantes

Entre estas variables, destaca el país del equipo local: la **variable Bolivia como equipo anfitrión** presenta el coeficiente más negativo de todo el modelo logístico y aparece entre las variables más importantes en todos los modelos. Esto se explica por su bajísima incidencia de lesiones durante el periodo analizado, especialmente en 2022 (ver Tabla 5).

La **distancia recorrida por el visitante** y **la altitud del partido** también aparecen de forma consistente como variables relevantes en los tres modelos. Estos factores, que reflejan condiciones logísticas y fisiológicas exigentes, contribuyen a predecir con mayor precisión el riesgo de lesión, especialmente tras el cambio de calendario en 2023. Por su parte, temperatura, aunque menos destacada en regresión logística, gana relevancia en modelos no lineales como XGBoost y Random Forest, lo que sugiere que su influencia puede depender de interacciones más complejas con otras variables.

Finalmente, las variables **de fases de clasificación**, que incluye la fase de playoffs y fases de octavos de final también tienen su importancia en el modelo de regresión logística, confirmando que el cambio de calendario de 2023 tiene su efecto en el riesgo de lesiones.

7.3 HERRAMIENTA INTERACTIVA DE PREDICCIÓN

Con el objetivo de facilitar el uso práctico del modelo desarrollado, se han implementado dos interfaces interactivas. La primera permite al usuario seleccionar las variables a incluir en el entrenamiento, así como filtrar el tipo de predicción deseada (lesión general, diagnóstico o localización). Da flexibilidad para explorar cómo cambia el rendimiento del modelo según se incorporen o excluyan factores contextuales (Figura 48 y Figura 49).

Figura 48: Interfaz de la herramienta de entrenamiento de modelos

```
=== MÉTRICAS STACKING ===
Accuracy:  0.586207
Precision:  0.434783
Recall:  0.769231
AUC ROC:  0.625375
F1-score:  0.555556

=== CLASSIFICATION REPORT ===
              precision    recall  f1-score   support

   No lesión       0.81      0.49      0.61        77
      Lesión       0.43      0.77      0.56        39

    accuracy                           0.59       116
   macro avg       0.62      0.63      0.58       116
weighted avg       0.68      0.59      0.59       116
```

Figura 49: Ejemplo de la herramienta de entrenamiento de modelo

La segunda interfaz permite introducir las características de un partido específico y obtener la probabilidad de lesión estimada por cada modelo, lo que puede ser útil para tomar decisiones preventivas en planificación de entrenamientos, rotaciones o cargas de trabajo (Figura 50 y Figura 51).

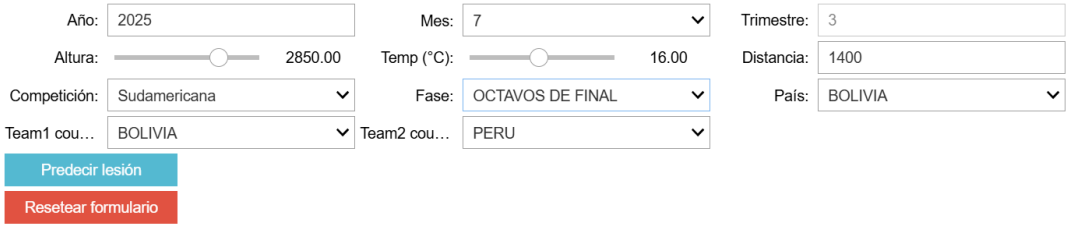

Figura 50: Interfaz de la herramienta de predicción de lesión en un partido

```
📋 Datos del partido:
 - Año: 2025
 - Mes: 7
 - Trimestre: 3
 - Altura (m): 2850.0
 - Temperatura (°C): 16.0
 - Distancia visitante (km): 1400.0
 - Días transcurridos desde 1 de enero: 181
 - Competición: Libertadores
 - Fase del torneo: OCTAVOS DE FINAL
 - País del partido: BOLIVIA
 - País del equipo local (Team1): BOLIVIA
 - País del equipo visitante (Team2): PERU

Entrenando modelo de Regresión Logística
Entrenando modelo de XGBoost
Entrenando modelo de Random Forest
Entrenando modelo de Stacking

🔎 Modelo: Regresión Logística
Predicción: ✅ Sin lesión
Probabilidad de lesión: 11.62%

🔎 Modelo: XGBoost
Predicción: ✅ Sin lesión
Probabilidad de lesión: 29.05%

🔎 Modelo: Random Forest
Predicción: ✅ Sin lesión
Probabilidad de lesión: 18.71%

🔎 Modelo: Stacking
Predicción: ✅ Sin lesión
Probabilidad de lesión: 46.54%
```

Figura 51: Ejemplo de resultado de salida de la herramienta de predicción de lesiones

Capítulo 8. CONCLUSIONES Y TRABAJOS FUTUROS

A lo largo de este trabajo se han alcanzado los objetivos planteados, tanto generales como específicos. En primer lugar, el análisis exploratorio del punto 5.4, ha permitido **identificar patrones significativos en relación con la aparición de lesiones**, destacando la influencia de variables como el tipo de competición, el año, la condición de local o visitante, y el país anfitrión. Estos resultados se recogen en la Tabla 9.

Factor analizado	Conclusión principal
Competición	La Copa Sudamericana presenta más lesiones que la Libertadores, especialmente en 2023 y 2024.
Año	En 2023 y 2024 aumentan las lesiones por la introducción de los playoffs y el calendario más comprimido.
Condición del equipo (local vs visitante)	Los jugadores locales sufren más lesiones que los visitantes, posiblemente por el sobreesfuerzo ante su afición y las faltas provocadas por el rival.
Altura y temperatura	No se encuentra relación lineal entre altitud o temperatura y las lesiones, aunque podrían influir combinadas con otras variables.
Distancia recorrida por el visitante	Desde el cambio de calendario en 2023, cuanto mayor es la distancia recorrida por el visitante, más lesiones se registran en el equipo local.
País anfitrión	A nivel geográfico, Bolivia presenta la menor incidencia, mientras que Colombia y Venezuela destacan por tasas altas, asociadas al mal estado de sus estadios y recientes reformas, respectivamente.

Tabla 9: Conclusiones del análisis exploratorio

Además los resultados obtenidos de los modelos predictivos principales (regresión logística, XGBoost y Random Forest), reflejados en la Tabla 8, confirman que las variables que mostraron patrones claros en el análisis exploratorio son también relevantes a nivel predictivo por lo que los modelos respaldan la hipótesis de que aspectos contextuales, como el país donde se juega, la distancia del desplazamiento o las condiciones ambientales, tienen un impacto directo sobre el riesgo de lesión.

Estas conclusiones permiten afirmar que se han cumplido los **objetivos generales** del proyecto: se ha llevado a cabo un análisis exploratorio detallado y se han evaluado variables contextuales clave, se han desarrollado modelos predictivos explicables capaces de estimar la probabilidad de lesión en función del contexto del partido. También se han alcanzado los **objetivos específicos,** como identificar variables individuales con gran peso explicativo, por ejemplo, el país del equipo anfitrión o la distancia recorrida por el visitante, analizar el efecto del calendario competitivo (especialmente tras la introducción de los playoffs en 2023), y realizar análisis por subgrupos según tipo de lesión o localización anatómica, todo ello integrado en la interfaz interactiva desarrollada.

Además, se ha construido una **herramienta práctica adaptable** que, con datos más avanzados o modelos perfeccionados en futuros trabajos, podría ser de utilidad para cuerpos técnicos y equipos médicos. Esta interfaz permite cargar datos de un partido y obtener una estimación del riesgo de lesión, lo que facilita la toma de decisiones preventivas. Su estructura escalable y flexible permite integrarla fácilmente en otros contextos, competiciones o niveles de detalle, cerrando así el ciclo entre análisis estadístico, modelado predictivo y aplicación práctica.

De cara a **trabajos futuros**, existen múltiples vías para continuar y ampliar el presente estudio. En primer lugar, se han identificado limitaciones importantes relacionadas con la información disponible: variables como los kilómetros exactos recorridos por cada jugador, el número de minutos disputados antes de la lesión o si esta fue provocada por contacto o surgió de forma espontánea, no están presentes en el dataset. En futuros trabajos, podrían incorporarse mediante técnicas como el análisis de vídeo o la visión por computador, que permitirían capturar automáticamente datos del rendimiento y los movimientos de los jugadores en el momento de la lesión. Esto permitiría, por ejemplo, asociar cada lesión a una carga de trabajo previa concreta o al tipo de interacción que la provocó.

También se podrían incluir variables contextuales adicionales para mejorar el análisis. Por ejemplo, una estimación del porcentaje de afición local frente a visitante en el estadio podría ofrecer información relevante, ya que como se ha observado, los equipos locales parecen

más propensos a lesionarse, posiblemente por presión ambiental. De forma similar, incorporar variables meteorológicas (lluvia, temperatura real durante el partido, humedad), tensión acumulada por resultados recientes o si el encuentro es amistoso u oficial, permitiría capturar condiciones externas que pueden aumentar el riesgo de lesión. Sin embargo, muchas de estas variables no están disponibles de forma directa y su incorporación requeriría una estrategia específica de recogida de datos, como acceder a APIs deportivas, procesar fuentes periodísticas o construir bases de datos manuales para casos concretos.

Desde el punto de vista de **modelado**, sería interesante evaluar el rendimiento del sistema con un mayor número de muestras para cada tipo específico de lesión o localización, ya que muchas categorías presentan muy pocos casos. Esto permitiría entrenar clasificadores individuales por tipo con mayor precisión.

Finalmente, han quedado fuera del alcance de este trabajo algunas extensiones: predecir no solo si habrá lesión, sino cuántas y en qué momento del partido pueden ocurrir; aplicar el análisis en otras competiciones, como ligas nacionales o divisiones inferiores; o replicar el enfoque en otras regiones geográficas como Europa o Asia para evaluar si las variables relevantes se mantienen.

Capítulo 9. BIBLIOGRAFÍA

[1] Barça Innovation Hub, «¿Cuánto se lesiona un jugador profesional? Un análisis de la epidemiología de las lesiones en el fútbol,» 31 Mayo 2021. [En línea]. Available: https://barcainnovationhub.fcbarcelona.com/es/blog/cuanto-se-lesiona-un-jugador-profesional-un-analisis-de-la-epidemiologia-de-las-lesiones-en-el-futbol/?.

[2] D. TVC, «Conoce famosos futbolistas que acabaron su carrera por lesiones,» 28 Junio 2023. [En línea]. Available: https://www.deportestvc.com/futbol-internacional/conoce-famosos-futbolistas-acabaron-carrera-lesiones-2023-06-28.

[3] E. Deportes, «Futbolistas, traspasos millonarios que no rindieron por lesiones,» 4 Enero 2021. [En línea]. Available: https://www.emol.com/noticias/Deportes/2021/01/04/1008404/Futbolistas-Traspasos-Millonarios.html?.

[4] Howden group, «Índice de lesiones de fútbol europeo masculino de Howden para 2023/24.,» 15 Octubre 2024. [En línea]. Available: https://www.howdengroup.com/es-es/reports/indice-de-lesiones-en-el-futbol-europeo-masculino-2023-24.

[5] A. M. Valencia, «Los peligros de que el fútbol europeo ficha a tantos latinoamericanos,» 13 Agosto 2014. [En línea]. Available: https://www.bbc.com/mundo/noticias/2014/08/140812_futbol_sudamerica_europa_amv?utm_source=chatgpt.com.

[6] Instituto de Biomecánica de Valencia (IBV), «IBV combina biomecánica e inteligencia artificial para desarrollar solcuiones innovadoras aplicables en el ámbito

de la salud y el deporte,» 3 Febrero 2025. [En línea]. Available: https://business-school.laliga.com/noticias/como-prevenir-lesiones-con-la-ayuda-de-la-inteligencia-artifical-ia.

[7] LaLiga Business School, «Cómo prevenir lesiones con la ayuda de la Inteligencia Artifical (IA),» 3 Febrero 2025. [En línea]. Available: https://business-school.laliga.com/noticias/como-prevenir-lesiones-con-la-ayuda-de-la-inteligencia-artifical-ia.

[8] Catapult Sports, «Prevención de lesiones en el deporte: Las ventajas de los sistemas de seguimiento de deportistas,» 5 Junio 2024. [En línea]. Available: https://www.catapult.com/es/blog/prevencion-de-lesiones-en-el-deporte.

[9] T. L., «Impacto de la altitud y el calor sobre el rendimiento en el fútbol,» 2014. [En línea]. Available: https://www.gssiweb.org/latam/sports-science-exchange/art%C3%ADculo/sse-131-impacto-de-la-altitud-y-el-calor-sobre-el-rendimiento-en-el-futbol?.

[10] «Pandas Documentation,» NumFocus, Inc., 5 Junio 2025. [En línea]. Available: https://pandas.pydata.org/docs/getting_started/index.html.

[11] C. R. Harris, K. J. Millman, S. J. Van der Walt, R. Gommers, P. Virtanen y et al., «NumPy,» *Nature,* vol. 585, nº 7825, pp. 357-362, Septiembre 2020.

[12] Python Software Foundation, «Datetime — Basic date and time type,» 2001-2025. [En línea]. Available: https://docs.python.org/es/3/library/datetime.html.

[13] Python Software Foundation, «Re — Regular expression operations,» 2001-2025. [En línea]. Available: https://docs.python.org/es/3/library/re.html.

[14] Python Software Foundation, «openpyxl 3.1.5,» 28 Junio 2024. [En línea]. Available: https://pypi.org/project/openpyxl/.

[15] J. McNamara, «Creating Excel files with Python and XlsxWriter,» 2013-2025. [En línea].

[16] J. Hunter, D. Darren, E. Firing, M. Droettboom y Matplotlib development team, «Matplotlib: Visualization with Python,» 2012-2025. [En línea]. Available: https://matplotlib.org/.

[17] M. Waskom, «Seaborn: statistical data visualization,» 2012-2024. [En línea]. Available: https://seaborn.pydata.org/.

[18] Scikit-learn developers, «Scikit-learn,» 2025. [En línea]. Available: https://scikit-learn.org/stable/supervised_learning.html.

[19] xgboost developers, «XGBoost Documentation,» 2022. [En línea]. Available: https://xgboost.readthedocs.io/en/stable/.

[20] The imbalanced-learn developers, «içImbalanced-learn documentation,» 20 December 2024. [En línea]. Available: https://imbalanced-learn.org/stable/.

[21] Python Software Foundation, «ipywidgets 8.1.7,» 5 Mayo 2025. [En línea]. Available: https://pypi.org/project/ipywidgets/.

[22] Catapult Sports, «Fundamentos del PlayerLoad™: cómo medir la carga de trabajo de un atleta,» 2024 Enero 2024. [En línea]. Available: https://www.catapult.com/es/blog/fundamentos-playerload-atleta-trabajo.

[23] Oliver Sports, «Cómo prevenir lesiones en el fútbol con tecnología GPS,» [En línea]. Available: https://blog.oliversports.ai/es/como-prevenir-lesiones-con-tecnologia-gps-en-el-futbol.

[24] IDE Universidad, «La biomecánica y la prevención de lesiones,» [En línea]. Available: https://ideuniversidad.com/enrique-navarro-cabello-la-biomecanica-y-la-prevencion-de-lesiones/.

[25] D. A. G. Bernal, «Los IMUs: el revolucionario avance de Podoactiva para el estudio de la pisada,» 27 Diciembre 2023. [En línea]. Available: https://www.podoactiva.com/blog/los-imus-el-revolucionario-avance-de-podoactiva-para-el-estudio-de-la-pisada.

[26] Cadena SER, «Podoactiva renueva como proveedor de podología y biomecánica del Real Valladolid,» 13 Septiembre 2024. [En línea]. Available: https://cadenaser.com/aragon/2024/09/13/podoactiva-renueva-como-proveedor-de-podologia-y-biomecanica-del-real-valladolid-radio-huesca/?.

[27] F. Bartels, L. Xing, C. Midoglu, M. Boeker, T. Kirsten y H. Pal, «SoccerGuard: Investigating Injury Risk Factors for Professional Soccer Players with Machine Learning,» 29 Octubre 2024. [En línea]. Available: https://arxiv.org/abs/2411.08901.

[28] AS, «El Marbella ficha a Olocip para optimizar la gestión de sus datos de salud y lesiones,» 18 Septiembre 2024. [En línea]. Available: https://as.com/futbol/mas_futbol/el-marbella-ficha-a-olocip-para-optimizar-la-gestion-de-sus-datos-de-salud-y-lesiones-n/.

[29] Innovación Digital 360, «Tecnología en el fútbol: Cómo el análisis predictivo y la IA transforman las prácticas del Manchester City,» 4 Septiembre 2024. [En línea]. Available: https://www.innovaciondigital360.com/i-a/tecnologia-en-el-futbol-como-

el-analisis-predictivo-y-la-inteligencia-artificial-transforman-los-entrenamientos-del-manchester-city/.

[30] A. d. Estal, «Prevención de lesiones con termografía en fútbol: resultados recientes,» 22 Junio 2022. [En línea]. Available: https://thermohuman.com/es/2022/06/23/prevencion-de-lesiones-con-termografia-en-futbol-resultados-recientes.

[31] LaLiga Business School, «LaLiga marca el camino de futuro del BI y Analytics en el fútbol gracias a Mediacoach y el proyecto Beyond Stats,» 1 Julio 2022. [En línea]. Available: https://www.laliga.com/noticias/laliga-marca-el-camino-de-futuro-del-bi-y-analytics-en-el-futbol-gracias-a-mediacoach-y-el-proyecto-beyond-stats.

[32] Glasdoor, «Sueldo de Ingeniero Software Junio. Glasdoor,» 2025. [En línea]. Available: https://www.glassdoor.es/Sueldos/ingeniero-de-software-junior-sueldo-SRCH_KO0,28.htm.

[33] FIFPRO, «WFS: ligas y sindicatos respaldan a los futbolistas por el calendario y reiteran su compromiso de emprender acciones legales,» 19 septiembre 2024. [En línea]. Available: https://fifpro.org/es/apoyar-a-los-y-las-futbolistas/salud-y-rendimiento/carga-de-trabajo-del-futbolista/wfs-ligas-y-sindicatos-respaldan-a-los-futbolistas-por-el-calendario-y-reiteran-su-compromiso-de-emprender-acciones-legales.

[34] L. Taylor y I. Rollo, «Impact of altitude and heat on football performance . Gatorade Sports Science Institute,» June 2014. [En línea]. Available: https://www.gssiweb.org/sports-science-exchange/article/sse-131-impact-of-altitude-and-heat-on-football-performance.

[35] Wikipedia Contributors, «Copa Libertadores de América. Wikipedia,» 2025. [En línea]. Available:

https://es.wikipedia.org/wiki/Copa_Libertadores_de_Am%C3%A9rica. [Último acceso: 27 Marzo 2025].

[36] Wikipedia contributors, «Copa Sudamericana. Wikipedia,» 2025. [En línea]. Available: https://es.wikipedia.org/wiki/Copa_Sudamericana. [Último acceso: 27 Marzo 2025].

[37] Wikipedia Contributors, «2024 Copa Sudamericana. Wikipedia,» 2024. [En línea]. Available: https://en.wikipedia.org/wiki/2024_Copa_Sudamericana. [Último acceso: 27 Marzo 2025].

[38] Conmebol, «Con cambios en el formato, la CONMEBOL Sudamericana gana aún más competitividad y atractivo.,» 22 Diciembre 2022. [En línea]. Available: https://www.conmebol.com/noticias/con-cambios-en-el-formato-la-conmebol-sudamericana-gana-aun-mas-competitividad-y-atractivo/.

[39] R. F. Chapman, T. Karlsen, G. k. Resaland, R.-L. Ge, M. P. Harber, S. Witowski, J. Stray-Gundersen y B. D. Levine, «Defining the 'dose' of altitude training: how high to live for optimal sea level performance enhancement,» *Journal of Applied Physiology,* pp. 595-603, 2014.

[40] J. Lopesino, «Estêvão, tortura a 3.650 metros,» 25 abril 2025. [En línea]. Available: https://as.com/futbol/internacional/estevao-se-exhibe-en-la-altura-n/.

[41] D. Link y H. Weber, «Effect of Ambient Temperature on Pacing in Soccer Depends on Skill Level,» *Journal of Strength and Conditioning Research,* pp. 1766-1770, Julio 2017.

[42] B. Turianski, «024 Copa Libertadores: Location-map for the 47-team tournament with Club Histories (Libertadores appearances & Titles listed),» 1 Febrero 2024. [En línea]. Available: https://billsportsmaps.com/wp-

content/uploads/2025/02/conmebol_copa-libertadores_2024_location-map_47-teams_m_.gif.

[43] J. f. García, «¿Cuál es la ciudad con más estadios de fútbol en el mundo?,» *WIN Sports,* 2 Septiembre 2024.

[44] S. C. Jasper, M. A. A. M. Leenders y T. O'Shannassy, «Travel across time zones and the implications for human performance post pandemic: Insights from elite sport,» *Front Public Health,* 22 Diciembre 2022.

[45] C. D. Rodríguez, «São Paulo vs. Independiente del Valle, resultado, resumen y goles: los ecuatorianos ganaron en Córdoba y se quedaron con la segunda Copa Sudamericana de su historia,» *The Sporting News,* 2 Octubre 2022.

[46] Wikipedia Contributors, «2023 Copa Sudamericana final.,» 29 Octubre 2023. [En línea]. Available: https://en.wikipedia.org/wiki/2023_Copa_Sudamericana_final.

[47] F. L. Zalcman, «Campeonato Brasileiro 2024: calendário completo e atualizado do torneio,» 26 Noviembre 2024. [En línea]. Available: https://www.olympics.com/pt/noticias/campeonato-brasileiro-2024-calendario-completo-atualizado.

[48] B. Scotti, «Liga Profesional Argentina 2024: ¿Cuándo empieza el Torneo y cuáles son las diferencias en el formato con la Copa de la Liga?,» 6 mayo 2024. [En línea]. Available: https://www.olympics.com/es/noticias/liga-profesional-argentina-2024-fecha-formato.

[49] D. S. Saha, «Analysis of time and injury incidences in I-league,» *International Journal of Physical Education, Sports and Health,* pp. 280-281, 2015.

[50] BeSoccer, «BeSoccer, Live Scores,» 2012-2025. [En línea]. Available: https://es.besoccer.com/competicion/info/copa_sudamericana.

[51] F. Wunderlich, M. Weigett, R. Rein y D. Memmert, «How does spectator presence affect football? Home advantage remains in European top-class football matches played without spectators during the COVID-19 pandemic,» *PLOS ONE,* 2021.

[52] D. Rodríguez, «Prueba deindependencia de Chi-cuadrado. Analytics Lane,» 6 mayo 2020. [En línea]. Available: https://www.analyticslane.com/2020/05/06/prueba-de-independencia-de-chi-cuadrado/.

[53] E. Soteldo, «Qué estadios de Venezuela recibieron inversión de Conmebol,» *El Impulso,* 18 Abril 2024.

[54] D. Rey, «Los stadios de Colombia están peor que los de la cuarta división de España, aseguró volante internacional del América de Cali,» *infobae,* 24 ferebro 2023.

[55] C. A. G. Berrum, «Una breve historia del machine learning: La tecnoogía que está cambiando nuestras vidas,» *Expost. Iexe Universidad,* 2022.

[56] N. Selvaraj, «8 modelos de machine learning explicados en 20 minutos. Datacamp,» 25 abril 2024. [En línea]. Available: https://www.datacamp.com/es/blog/machine-learning-models-explained.

[57] F. J. Reis, R. K. Alaiti, C. Sain Vallio y L. Hespanhol, «Artificial intelligence and Machine Learning approaches in sports: Concepts, applications, challenges, and future perspectives,» *PubMed Centrl,* 2024.

[58] Alvro, «Inteligencia artificial para la detección del cáncer. Machine Learning para todos,» 6 julio 2020. [En línea]. Available: https://machinelearningparatodos.com/inteligencia-artificial-deteccion-cancer/.

[59] A. Munoz-MAcho, M. Dominguez-Morales y J. Sevillano-Ramos, «Performance and healthcare analysis in elite sports teams using artificial intelligence: a scoping review,» *Front Sports Act Living.,* 2024.

[60] J. Murel y E. Kavlakoglu, «¿Qué son los modelos de clasificación?. IBM,» 31 julio 2024. [En línea]. Available: https://www.ibm.com/es-es/think/topics/classification-models#:~:text=Los%20modelos%20de%20clasificaci%C3%B3n%20son%20un%20tipo%20de,puntos%20de%20datos%20en%20grupos%20predefinidos%20denominados%20clases..

[61] N. Selvaraj, «8 modelos de machine learning explicados en 20 minutos. Datacamp,» 25 Abril 2024. [En línea]. Available: https://www.datacamp.com/es/blog/machine-learning-models-explained.

[62] J. J. E. Zúñiga, «Aplicación de algoritmos Random Forest y XGBoost en una base de solicitudes de tarjetas de crédito,» *Ingeniería Investigación y Tecnología,* pp. 1-16, 2020.

[63] H. Saleh, The Machine Learning Workshop, Birmingham, UK: Packt Publishing, 2020.

[64] Scikit-learn developers, «Feature Selection. Scikit Learn,» 2025. [En línea]. Available: https://scikit-learn.org/stable/modules/feature_selection.html.

[65] N. Unidas, «Objetivo 3: Garantizar una vida sana y promover el bienestar para todos en todas las edades,» [En línea]. Available: https://sdgs.un.org/es/goals/goal3.

[66] N. Unidas, «Objetivo 9: Construir infraestructuras resilientes, promover la industrialización inclusiva y sostenible y fomentar la innovación,» s.f. [En línea]. Available: https://sdgs.un.org/es/goals/goal9.

[67] N. Unidas, «Objetivo 12: Garantizar modalidades de consumo y producción sostenibles,» [En línea]. Available: https://sdgs.un.org/es/goals/goal12.

[68] A. P. L. C. P. I. F. M. F. J. &. M. D. Rossi, «Effective injury forecasting in soccer with GPS training data and machine learning,» *Arxiv,* pp. 1-40, 2017.

[69] OpenAI, «ChatGPT (GPT-4o),» 2023. [En línea]. Available: https://chat.openai.com/.

ANEXO I: ALINEACIÓN DEL PROYECTO CON LOS ODS

Este Trabajo Fin de Grado guarda un alineamiento claro con varios de los Objetivos de Desarrollo Sostenible (ODS) propuestos por la Organización de las Naciones Unidas, especialmente en lo relativo a la salud, la innovación tecnológica y la eficiencia en el uso de recursos en el ámbito deportivo. A través del análisis predictivo y contextualizado de lesiones en el fútbol sudamericano, el proyecto contribuye a mejorar las condiciones físicas y competitivas de los deportistas, a impulsar el uso de tecnología en sectores no siempre priorizados, y a fomentar una gestión deportiva más racional, basada en datos.

ODS 3: Salud y Bienestar: "Garantizar una vida sana y promover el bienestar para todos en todas las edades"

El objetivo principal de este proyecto es prevenir lesiones deportivas a través del análisis de datos y el uso de modelos predictivos. Esto se traduce en una mejora directa del bienestar físico y mental de los jugadores, al evitar bajas médicas, recaídas y consecuencias psicológicas derivadas del sobreesfuerzo o la falta de descanso. Además, la propuesta está especialmente adaptada a las condiciones del fútbol sudamericano, donde los recursos médicos y de prevención pueden ser limitados. En ese sentido, el desarrollo de soluciones interpretables, basadas en datos accesibles y sin requerimientos tecnológicos complejos, permite democratizar el acceso a herramientas de salud preventiva en contextos menos favorecidos [65].

ODS 9: Industria, Innovación e Infraestructura: "Construir infraestructuras resilientes, promover la industrialización inclusiva y sostenible y fomentar la innovación"

Este proyecto utiliza tecnologías avanzadas como el procesamiento de datos masivos, la inteligencia artificial y los modelos de machine learning, adaptadas al entorno deportivo. Se promueve así una innovación digital accesible, especialmente útil para ligas o clubes que no pueden permitirse sistemas comerciales costosos como los utilizados en el fútbol europeo.

El enfoque combina variables contextuales (altitud, acumulación de partidos, condición de local o visitante) con datos de lesiones, generando una infraestructura analítica que fortalece la toma de decisiones en cuerpos técnicos y médicos. Esto permite avanzar hacia una gestión deportiva más tecnológica, eficiente y basada en evidencia [66].

ODS 12: Producción y Consumo Responsables: "Garantizar modalidades de consumo y producción sostenibles"

El proyecto también se alinea con la sostenibilidad en el uso de recursos humanos dentro del deporte. Las estrategias propuestas están orientadas a optimizar la planificación física de los jugadores, evitando sobrecargas innecesarias, distribuyendo mejor los esfuerzos a lo largo de la temporada y facilitando decisiones informadas en función del contexto.
Esta eficiencia contribuye a un uso responsable del “recurso jugador”, maximizando su rendimiento a lo largo del tiempo y reduciendo los impactos negativos derivados de una mala gestión de las cargas físicas o de decisiones tácticas que no consideran el estado real del futbolista. En contextos como el sudamericano, donde la plantilla es limitada y los recursos médicos escasos, esta racionalización tiene un valor especialmente relevante [67].

ANEXO II: CÓDIGOS DE DIAGNÓSTICO Y LOCALIZACIÓN

1. Códigos de diagnóstico

1: FRACTURA – Rotura de hueso.

3: HERIDA – Corte, laceración u otra lesión abierta.

4: MUSCULAR – Dolor o rigidez no específica.

5: ESGUINCE – Lesión de ligamentos, generalmente leve.

6: CONTUSIÓN – Golpe sin rotura, común en deportes de contacto.

7: LUMBALGIA – Dolor en la zona lumbar de la espalda.

8: LCA – Rotura del ligamento cruzado anterior.

9: FASCITIS – Inflamación de tejido blando (ej. fascitis plantar).

10: LUXACIÓN – Dislocación de una articulación.

11: MENISCO – Lesión del cartílago meniscal de la rodilla.

12: TENDINITIS – Inflamación de tendón.

13: PARADA CARDIORRESPIRATORIA – Evento médico severo.

14: GOLPE DE CALOR, VÓMITOS – Síntomas de estrés térmico o agotamiento.

20: CONCUSIÓN LEVE – Golpe leve en la cabeza con síntomas menores.

21: CONCUSIÓN GRAVE – Lesión cerebral traumática severa.

22: CONCUSIÓN + HERIDA FACIAL – Doble afectación por golpe.

24: VÓMITOS – Posiblemente por esfuerzo, calor o golpe.

41: CONTRACTURA – Acortamiento o tensión muscular dolorosa.

42: DESGARRO – Ruptura de fibras musculares.

43: AVULSIÓN – Desinserción de un tendón o ligamento.

44: CONTUSIÓN – (Repetido del código 6) Daño por golpe.

45: ROTURA – Rotura completa de músculo o tendón.

46: CALAMBRE – Contracción muscular involuntaria.

51: ESGUINCE RODILLA LLI – Lesión del ligamento lateral interno.

52: ESGUINCE TOBILLO – Muy común en deportes de contacto.

53: ESGUINCE RODILLA LLE – Lesión del ligamento lateral externo.

81: LCA + LCP + MENISCO – Lesión combinada de ligamentos y cartílago.

82: LCA + MENISCO + LCP – Igual que 81, distinto orden.

101: SUBLUXACIÓN – Dislocación parcial de una articulación.

2. Códigos de localización

1: CABEZA – Parte superior del cuerpo, puede implicar conmociones o cortes.

2: CARA – Lesiones faciales como fracturas nasales o cortes.

3: CUELLO – Incluye lesiones cervicales o contracturas.

4: HOMBRO – Luxaciones, contusiones o roturas musculares.

5: BRAZO – Lesiones de bíceps o huesos del brazo.

6: CODO – Luxaciones, epicondilitis u otras lesiones articulares.

7: ANTEBRAZO – Incluye fracturas de radio/cúbito o contusiones.

8: MUÑECA – Lesiones articulares o tendinosas.

9: MANO – Incluye dedos y zona palmar.

10: DEDOS MANO – Lesiones específicas en los dedos.

11: TÓRAX – Contusiones costales, fracturas o traumatismos.

12: COLUMNA TORÁCICA – Lesiones vertebrales en esa región.

13: COLUMNA LUMBAR – Contracturas o lumbalgias.

14: SACRO – Lesiones en la base de la columna.

15: PELVIS – Fracturas, contusiones o problemas articulares.

16: CADERA – Incluye tendinitis o luxaciones.

17: FÉMUR – Fracturas o contusiones en el hueso del muslo.

18: MUSLO – Lesiones musculares como desgarros o hematomas.

80: MÚSCULO CUÁDRICEPS – Desgarros o hematomas.

181: ISQUIOTIBIALES – Contracturas o roturas.

182: BÍCEPS FEMORAL – Desgarros, contracturas o tendinopatías.

183: ADUCTORES – Frecuente en sobrecargas o estiramientos forzados.

19: RODILLA – Lesiones ligamentarias o meniscales.

20: PIERNA – Incluye tibia/peroné y parte inferior.

23: DEDO PIE – Lesiones puntuales como fracturas o esguinces.

24: GENERAL – Cuando no se puede especificar claramente.