A picture containing text, font, screenshot, graphics

Description automatically generated

**Facultat Internacional de Comerç i Economia Digital La Salle**

Trabajo Final de Máster

Máster Universitario en Ciencia de los Datos / Data Science

TÍTULO TFM

Alumno Profesor Ponente

Bernat Sort Rufat Roger Mallol Parera

|  |
| --- |
| **ACTA DEL EXAMEN**  **DEL TRABAJO FINAL DE MÁSTER** |

Reunido el Tribunal calificador en el día de la fecha, el alumno

**Bernat Sort Rufat**

Expuso su Trabajo de Final de Máster, el cual trató sobre el tema siguiente:

**TÍTULO DEL TFM**

Acabada la exposición y contestadas por parte del alumno las objeciones formuladas por los miembros del tribunal, este valoró el mencionado Trabajo con la calificación de

|  |
| --- |
|  |

Barcelona,

VOCAL DEL TRIBUNAL VOCAL DEL TRIBUNAL

PRESIDENTE DEL TRIBUNAL

Agradecimientos

Resumen

Abstract

Tabla de Contenidos

[1 Introducción 1](#_Toc143968982)

[1.1 Contexto 1](#_Toc143968983)

[1.2 Motivación y justificación 2](#_Toc143968984)

[1.3 Objetivos 2](#_Toc143968985)

[1.4 Hipótesis 2](#_Toc143968986)

[1.5 Estructura del documento 3](#_Toc143968987)

[2 Estado del arte 4](#_Toc143968988)

[3 Marco teórico 7](#_Toc143968989)

[4 Métodos 7](#_Toc143968990)

[4.1 Herramientas y software 7](#_Toc143968991)

[4.2 El conjunto de datos GaitRec 8](#_Toc143968992)

[4.2.1 Introducción al conjunto de datos GaitRec 8](#_Toc143968993)

[4.2.2 Preparación y procesamiento 8](#_Toc143968994)

[4.2.3 Descripción y anotaciones 9](#_Toc143968995)

[4.2.4 Estructura de los datos 11](#_Toc143968996)

[4.3 Construcción de los conjuntos de datos 13](#_Toc143968997)

[4.4 Análisis y comprobaciones iniciales 19](#_Toc143968998)

[4.5 Estudio de la población y cohorte de pacientes 20](#_Toc143968999)

[4.6 Análisis exploratorio de los datos 21](#_Toc143969000)

[4.7 División de los datos 22](#_Toc143969001)

[4.8 Preprocesamiento de los datos 22](#_Toc143969002)

[4.8.1 Primera aproximación: datos sin procesar 22](#_Toc143969003)

[4.8.2 Segunda Aproximación: PCA en los datos de entrada (X) 22](#_Toc143969004)

[4.8.3 Tercera Aproximación: PCA en los datos de salida (Y) 23](#_Toc143969005)

[4.8.4 Cuarta aproximación: PCA en los datos de entrada y salida (X y Y) 23](#_Toc143969006)

[4.9 Análisis multivariante 24](#_Toc143969007)

[4.9.1 Tipo de problema y algoritmos utilizados 24](#_Toc143969008)

[4.9.2 Pipeline general de entrenamiento y evaluación de los modelos 24](#_Toc143969009)

[4.9.3 Regresión Linear Múltiple (MLR) 27](#_Toc143969010)

[5 Resultados 32](#_Toc143969011)

[5.1 EDA 38](#_Toc143969012)

[5.2 Regresión Linear Múltiple 39](#_Toc143969013)

[5.2.1 Primera aproximación 39](#_Toc143969014)

[6 Discusiones 40](#_Toc143969015)

[6.1 Limitaciones del estudio 40](#_Toc143969016)

[7 Consideraciones éticas 40](#_Toc143969017)

[8 Coste económico 40](#_Toc143969018)

[9 Conclusiones 40](#_Toc143969019)

[10 Vías abiertas 40](#_Toc143969020)

[11 Referencias 41](#_Toc143969021)

[12 Apéndice 43](#_Toc143969022)

[12.1 Apéndice 1. Repositorio del proyecto 43](#_Toc143969023)

Índice de Figuras

[Figure 1. Flowchart of this study. This flowchart shows the patient selection and the different steps applied in each cohort. **Error! Bookmark not defined.**](#_Toc110619725)

[Figure 2. Daily frequency of PaO2 records per patient versus the daily frequency of SpO2 records per patient. **Error! Bookmark not defined.**](#_Toc110619726)

[Figure 3. Correlation between PaO2/FiO2 and SpO2/FiO2 ratios. **Error! Bookmark not defined.**](#_Toc110619727)

[Figure 4. SpFi distribution based on each respiratory SOFA PaFi score. **Error! Bookmark not defined.**](#_Toc110619728)

[Figure 5. Groups (0, 1, 2, 3, and 4) of the new respiratory SOFA score based on the SpO2/FiO2 ratio. **Error! Bookmark not defined.**](#_Toc110619729)

[Figure 6. Histogram showing the distance between SOFA SpFi score and SOFA PaFi score. **Error! Bookmark not defined.**](#_Toc110619730)

[Figure 7. Comparison of PaFi and SpFi AUCs obtained by the several ML algorithms for classification. **Error! Bookmark not defined.**](#_Toc110619731)

Índice de Tablas

[Table 1. The criteria for assessment of the respiratory SOFA score [12]. **Error! Bookmark not defined.**](#_Toc110594726)

[Table 2. Variable selection. The criteria for assessment of the respiratory SOFA score [12]. **Error! Bookmark not defined.**](#_Toc110594727)

[Table 3. Features and target dataset example of PaFi on subset srpafi4% > 0\* **Error! Bookmark not defined.**](#_Toc110594728)

[Table 4. Features and target dataset example of SpFi on subset srpafi4% > 0\* **Error! Bookmark not defined.**](#_Toc110594729)

[Table 5. Performance metric for the machine learning algorithms in both datasets: AUC. **Error! Bookmark not defined.**](#_Toc110594730)

# Introducción

## Contexto

En los últimos años, el panorama de la investigación en el ámbito de la salud ha experimentado notables transformaciones debido a la creciente disponibilidad de grandes conjuntos de datos, provocado un creciente interés en la aplicación de métodos de aprendizaje automático y aprendizaje profundo [*Eric J Topol. 2019. High-performance medicine: The convergence of human and artificial intelligence. Nature Medicine 25, 1 (2019), 44–56. https://doi.org/10.1038/s41591-018-0300-7*].

Esta tendencia también se ha extendido en el campo del análisis clínico de la marcha (ACM) [*Joana Figueiredo, Cristina P. Santos, and Juan C. Moreno. 2018. Automatic recognition of gait patterns in human motor disorders using machine learning: A review. Medical Engineering and Physics 53 (2018), 1–12. https://doi.org/10.1016/j.medengphy.2017.12.006*],[*Wolfgang I Schöllhorn. 2004. Applications of artificial neural nets in clinical biomechanics. Clinical Biomechanics 19, 9 (2004), 876–898. https://doi.org/10.1016/j.clinbiomech.2004.04.005*]. Este campo interdisciplinar desempeña un papel fundamental en la comprensión de la biomecánica del movimiento humano, especialmente en personas con discapacidades musculoesqueléticas. El análisis clínico de la marcha tiene el potencial de ofrecer una visión profunda de las limitaciones funcionales y las alteraciones biomecánicas que experimentan estos pacientes, guiando así el desarrollo de intervenciones terapéuticas y estrategias de rehabilitación eficaces.

La biomecánica, el análisis de la marcha y la rehabilitación constituyen componentes integrales del panorama médico, cuyo objetivo es mejorar el bienestar y la capacidad funcional de las personas con deficiencias musculoesqueléticas. Los trastornos musculoesqueléticos abarcan una amplia gama de afecciones que afectan a huesos, músculos, ligamentos y articulaciones, y que a menudo provocan alteraciones en los patrones de movimiento y reducen la movilidad. Abordar estas deficiencias mediante terapias y regímenes de rehabilitación adaptados es fundamental no sólo para recuperar la función física, sino también para mejorar la calidad de vida general de las personas afectadas.

Sin embargo, hasta la fecha no se ha encontrado una forma de garantizar que la rehabilitación sea eficaz para lograr un cambio significativo en la vida del paciente, lo que a veces supone una grave pérdida de tiempo y dinero tanto para el paciente como para el sistema sanitario. Aunque la rehabilitación es muy prometedora a la hora de restablecer las funciones y mejorar el bienestar de las personas con deficiencias musculoesqueléticas, su eficacia varía mucho debido a la compleja interacción de factores que influyen en los resultados de la terapia. Esta imprevisibilidad supone un reto importante tanto para los médicos como para los pacientes, que invierten un tiempo, un esfuerzo y unos recursos considerables en programas de rehabilitación con resultados inciertos.

Es en este contexto donde la importancia de esta investigación emerge con mayor claridad. La laguna existente entre el impacto potencial de la rehabilitación y la realidad de unos resultados inciertos exige un cambio de paradigma. La integración de técnicas de aprendizaje automático para predecir el progreso de la terapia en pacientes con deficiencias musculoesqueléticas promete transformar el panorama de la rehabilitación aprovechando el valor de los datos.

## Motivación y justificación

En este contexto, la base de esta investigación se fundamente en una de las propuestas presentadas por Brian Horsak et al. en el artículo titulado *GaitRec: A Large-Scale Ground Reaction Force Dataset of Healthy and Impaired Gait* [gaitrec]. Este artículo proporciona un conjunto de datos exhaustivo y plantea cuestiones y áreas de investigación que se presentan como un terreno fértil para explorar y profundizar.

Esta investigación encuentra su motivación en la necesidad de reducir la brecha entre los objetivos de la rehabilitación y los resultados concretos alcanzados por individuos con deficiencias musculoesqueléticas. No obstante, las implicaciones de evaluar y predecir eficazmente el progreso de la terapia van más allá del ámbito clínico. Dotar a los médicos de herramientas para adaptar las intervenciones en función de indicadores objetivos de progreso es una innovación que puede redefinir el panorama del tratamiento. Los pacientes, equipados con una comprensión más clara de su proceso de rehabilitación, pueden embarcarse en el camino de la recuperación con un sentido de propósito y dirección. Además, la posibilidad de optimizar la asignación de recursos dentro de los sistemas sanitarios, junto con el potencial para mitigar el coste emocional, físico y económico de los pacientes, guía la dirección de esta investigación.

La incertidumbre que rodea la efectividad de las intervenciones de rehabilitación no solo plantea un desafío práctico, sino también una preocupación profundamente humanitaria. A medida que las personas emprenden el camino hacia la recuperación, invirtiendo esperanzas, energía y recursos, la incapacidad para anticipar la evolución del progreso puede resultar en desilusión y frustración.

Desde una perspectiva social, las ineficiencias derivadas de programas de rehabilitación ineficaces o mal adaptados pueden ejercer una carga excesiva sobre los sistemas de atención sanitaria. La motivación subyacente de este proyecto surge en respuesta a estas inquietudes, como un llamada a tomar medidas con el fin de trazar una ruta más sólida y fundamentada hacia el éxito en la rehabilitación.

La convergencia de la ciencia de datos y la biomecánica ofrece una oportunidad que resuena directamente con el núcleo de la motivación de esta investigación. Aprovechando el valor de los datos, este estudio busca elevar el proceso de toma de decisiones terapéuticas desde el ámbito de la intuición y la experiencia hacia la evidencia y el análisis riguroso. En última instancia, la finalidad es mejorar la calidad de vida de las personas afectadas por deficiencias musculoesqueléticas y contribuir al desarrollo de enfoques de rehabilitación más eficaces y fundamentados.

## Objetivos

El objetivo principal de este proyecto es analizar y predecir, mediante el desarrollo y la evaluación de distintos modelos de aprendizaje automático, el progreso y la evolución de los pacientes con deficiencias musculoesqueléticas en el tobillo sometidos a dos sesiones de terapia o rehabilitación.

## Hipótesis

Se parte de la hipótesis de que los modelos de aprendizaje automático desarrollados y entrenados utilizando un amplio conjunto de datos anotados, que contiene paseos o *trials* de marcha con fuerzas de reacción del suelo (GRF) bilaterales de pacientes que presentan diversas discapacidades musculoesqueléticas [gaitrec], demostrarán la capacidad de predecir de manera precisa el progreso y la evolución de los pacientes sometidos a terapia o rehabilitación.

## Estructura del documento

En la introducción se presenta el contexto necesario para la investigación y se sientan las bases de nuestro estudio. Se comienza introduciendo el campo de la biomecánica, el análisis de la marcha y la rehabilitación. Se discute la relevancia de comprender y evaluar los patrones de la marcha en individuos con deficiencias musculoesqueléticas y el impacto potencial en su calidad de vida. Se explican los retos que plantea la incertidumbre de los resultados de la rehabilitación y las posibles ineficiencias de los sistemas sanitarios. Se plantean los objetivos de nuestra investigación, así como hipótesis que guiará la investigación.

En el marco teórico, se profundiza en el conjunto de conocimientos existentes relacionados con nuestro tema de investigación. Se realiza una revisión de estudios similares relacionados con la evaluación de terapias e intervenciones de individuos con deficiencias musculoesqueléticas, así como de las aplicaciones del aprendizaje automático en la predicción del progreso y evolución de los pacientes. Finalmente, se destacan algunas de las limitaciones del panorama actual.

En el apartado de métodos, se presentan los procedimientos adoptados para la realización de la investigación. El primer apartado aborda la descripción exhaustiva del conjunto de datos empleado, incluyendo el número de participantes involucrados, los diversos tipos de deficiencias musculoesqueléticas abarcadas y el proceso riguroso de recolección de datos. A continuación, se presenta el análisis exploratorio de los datos, seguido de la exposición de los pasos de preprocesamiento de datos y la ingeniería de características o *feature engineering*. Posteriormente, se detalla el proceso de división de los datos para a la fase de entrenamiento, validación y pruebas. Se destaca además la elección y la justificación de los algoritmos de aprendizaje automático seleccionados. Para finalizar, se abordan en detalle las métricas de evaluación empleadas para valorar el desempeño y rendimiento de los modelos de aprendizaje automático.

En el apartado de resultados y discusión, se evalúa el rendimiento de los distintos modelos de aprendizaje automático utilizados empleando métricas de evaluación como la raíz del error cuadrático medio (RMSE) [] y el coeficiente de determinación (R2) []. Se compara los resultados de los distintos modelos y se seleccionan los que mejor rendimiento hayan tenido a la hora de predecir el progreso y la evolución de los pacientes con deficiencias musculoesqueléticas sometidos a terapia o rehabilitación. Se interpretan las implicaciones de los resultados en relación con la hipótesis y objetivos, se contextualizan los resultados, y se analiza su importancia en el ámbito de la evaluación del progreso de la rehabilitación.

En el siguiente apartado, se discuten las consideraciones éticas en torno al uso de los datos de los pacientes, haciendo hincapié en las medidas que se tomaron para garantizar la privacidad del paciente y la seguridad de los datos. Además, se reflexiona sobre el impacto potencial de nuestra investigación en la atención al paciente y las implicaciones éticas del uso de modelos predictivos en la toma de decisiones clínicas.

En la sección de coste económico, se analizan las implicaciones económicas relacionadas con nuestro proyecto de investigación. Se realiza una estimación de los recursos económicos requeridos para la ejecución del proyecto, abarcando tanto los costos de los materiales necesarios para llevar a cabo la investigación como el cálculo del salario por hora del científico de datos encargado de su realización. Esta evaluación económica proporciona una visión completa de los recursos financieros necesarios para el desarrollo y la ejecución exitosa del proyecto de investigación en cuestión.

En la sección de conclusiones, se presentan de manera concisa las conclusiones más destacadas y las contribuciones significativas derivadas de nuestra investigación. Se revisan los objetivos e hipótesis planteadas en el proyecto con el propósito de evaluar si se han alcanzado satisfactoriamente, y se destaca la importancia del trabajo realizado en el avance del campo de la evaluación del progreso de la rehabilitación y su posible impacto en la atención al paciente y los sistemas sanitarios.

Finalmente, en el apartado de vías abiertas, se proponen posibles vías de investigación y desarrollo en este campo. Se reflexiona cómo nuestro trabajo puede ampliarse, refinarse o adaptarse a diferentes contextos. También se aportan ideas sobre cómo podrían integrarse nuestros modelos predictivos en la práctica clínica, contribuyendo a la atención personalizada de los pacientes y a la optimización de las estrategias de rehabilitación.

# Estado del arte

En los últimos años, las técnicas de aprendizaje automático y aprendizaje profundo han demostrado una notable eficacia en el ámbito del análisis clínico de la marcha (ACM). Estos métodos han demostrado ser exitosos en la categorización de cohortes de pacientes [*Joana Figueiredo, Cristina P. Santos, and Juan C. Moreno. 2018. Automatic recognition of gait patterns in human motor disorders using machine learning: A review. Medical Engineering and Physics 53 (2018), 1–12.* [*https://doi.org/10.1016/j.medengphy.2017.12.006*](https://doi.org/10.1016/j.medengphy.2017.12.006)], [*Wolfgang I Schöllhorn. 2004. Applications of artificial neural nets in clinical biomechanics. Clinical Biomechanics 19, 9 (2004), 876–898. https://doi.org/10.1016/j.clinbiomech.2004.04.005*] abarcando condiciones como el ictus [*Hong-yin Lau, Kai-yu Tong, and Hailong Zhu. 2009. Support vector machine for classification of walking conditions of persons after stroke with dropped foot. Human Movement Science 28, 4 (Aug. 2009), 504–514. https://doi.org/10.1016/j.humov.2008.12.003*], casos de parálisis cerebral [*Leen Van Gestel, Tinne De Laet, Enrico Di Lello, Herman Bruyninckx, Guy Molenaers, Anja Van Campenhout, Erwin Aertbeliën, Mike Schwartz, Hans Wambacq, Paul De Cock, and Kaat Desloovere. 2011. Probabilistic gait classification in children with cerebral palsy: A Bayesian approach. Research in Developmental Disabilities 32, 6 (Nov. 2011), 2542–2552. https://doi.org/10.1016/j.ridd.2011.07.004*], casos de osteoartritis [*Corina Nüesch, Victor Valderrabano, Cora Huber, Vinzenz von Tscharner, and Geert Pagenstert. 2012. Gait patterns of asymmetric ankle osteoarthritis patients. Clinical Biomechanics 27, 6 (July 2012), 613–618. https://doi.org/10.1016/j.clinbiomech.2011.12.016*], esclerosis múltiple [*Murad Alaqtash, Thompson Sarkodie-Gyan, Huiying Yu, Olac Fuentes, Richard Brower, and Amr Abdelgawad. 2011. Automatic classification of pathological gait patterns using ground reaction forces and machine learning algorithms. In 2011 Annual International*

*Conference of the IEEE Engineering in Medicine and Biology Society (EMBS). IEEE, 453–457.* [*https://doi.org/10.1109/IEMBS.2011*](https://doi.org/10.1109/IEMBS.2011)*. 6090063*], Parkinson [*Ferdous Wahid, Rezaul K Begg, Chris J Hass, Saman Halgamuge, and David C Ackland. 2015. Classification of Parkinson’s disease gait using spatial-temporal gait features. IEEE Journal of Biomedical and Health Informatics 19, 6 (2015), 1794–1802.*], e individuos lidiando con diversos trastornos funcionales de la marcha [*Djordje Slijepcevic, Matthias Zeppelzauer, Anna-Maria Gorgas, Caterine Schwab, Michael Schüller, Arnold Baca, Christian Breiteneder, and Brian Horsak. 2017. Automatic classification of functional gait disorders. IEEE Journal of Biomedical and Health Informatics 22, 5 (2017), 1653–1661. https://doi.org/10.1109/JBHI.2017.2785682*].

En el ámbito del análisis predictivo de la marcha, los investigadores han explorado la aplicación de técnicas de aprendizaje automático para predecir los resultados postoperatorios en individuos con trastornos del movimiento. Además, en el ámbito de las intervenciones ortopédicas se han realizado contribuciones destinadas a mejorar la toma de decisiones y los resultados del tratamiento.

No obstante, a pesar de la proliferación de estudios que emplean técnicas de aprendizaje automático para analizar patrones y anomalías de la marcha, sigue existiendo un vacío a la hora de comprender y predecir la evolución de los pacientes con deficiencias musculoesqueléticas a lo largo de los procesos terapéuticos y de rehabilitación. Es por ello que, aunque los artículos en los que nos hemos basado para realizar nuestra investigación no están directamente relacionados con la evaluación y predicción del progreso de la terapia de la marcha, si presentan similitudes en cuanto al uso del aprendizaje automático para predecir resultados relacionados con los patrones de la marcha.

El artículo científico titulado *Predicting Postoperative Knee Flexion during Gait of Cerebral Palsy Children*, de Omar A. Galarraga C. et al. [O. A. Galarraga C., V. Vigneron, B. Dorizzi, N. Khouri & E. Desailly (2015) Predicting postoperative knee flexion during gait of cerebral palsy children, Computer Methods in Biomechanics and Biomedical Engineering, 18:sup1, 1940-1941, DOI: 10.1080/10255842.2015.1070583] se centra en el uso del aprendizaje automático para predecir la flexión postoperatoria de la rodilla durante la marcha en niños con parálisis cerebral. El estudio aborda los retos que plantean trastornos neurológicos como la parálisis cerebral, que afectan negativamente a la marcha y la movilidad. El estudio emplea el análisis de componentes principales (PCA) para reducir la dimensionalidad de las características de entrada. Utilizando un modelo de regresión lineal múltiple, los autores predicen la flexión de la rodilla durante la marcha postoperatoria, teniendo en cuenta los datos cinemáticos preoperatorios proyectados en un espacio de dimensión inferior mediante el análisis de componentes principales (PCA), los resultados de la exploración física y el tipo de cirugía. Mediante una cuidadosa consideración de varias dimensiones, los investigadores logran predicciones precisas de la flexión de la rodilla. La incorporación de un código de cirugía mejora la capacidad predictiva del modelo, que abarca varias categorías quirúrgicas y sus combinaciones. Los resultados se presentan con intervalos de confianza, lo que proporciona una medida fiable de la certeza de la predicción, y evalúan el rendimiento de la predicción mediante la raíz del error cuadrático medio (RMSE). Además, se destaca la posible utilidad clínica de los modelos predictivos para ayudar en la toma de decisiones sobre intervenciones quirúrgicas y ofrecer a los pacientes información sobre los resultados probables.

El artículo científico *Estimation of Postoperative Knee Flexion at Initial Contact of Cerebral Palsy Children using Neural Networks* de Omar A. Galarraga C. et al. [Galarraga, Omar & Vigneron, Vincent & Dorizzi, Bernadette & Khouri, Nejib & Desailly, Eric. (2015). Estimation of Postoperative Knee Flexion at Initial Contact of Cerebral Palsy Children using Neural Networks. ICPRAM 2015 - 4th International Conference on Pattern Recognition Applications and Methods, Proceedings. 2. 10.5220/0005286503380342.], profundiza en la estimación de la flexión postoperatoria de la rodilla, un parámetro crucial en el contexto de la parálisis cerebral y las intervenciones ortopédicas. El estudio aborda el reto de predecir la flexión de la rodilla tras la cirugía de alargamiento de los isquiotibiales, una intervención ortopédica habitual para aliviar dichas irregularidades de la marcha. Esta investigación aprovecha el potencial de las redes neuronales para el ajuste no lineal de datos, demostrando la viabilidad de predecir la flexión postoperatoria de la rodilla basándose en datos preoperatorios de la marcha, la exploración física y el tipo de cirugía. Los autores emplean una única arquitectura de red neuronal que toma múltiples variables de entrada relacionadas con el estado del paciente y la cirugía, y produce como salida los ángulos de flexión de rodilla predichos en el contacto inicial. La medida del error considerada para cada paciente es la raíz del error cuadrático medio (RMSE). Mediante técnicas de aprendizaje supervisado, el estudio establece un simulador capaz de estimar los resultados de la flexión postoperatoria de la rodilla con un encomiable nivel de precisión.

El artículo científico titulado *Predicting Postoperative Gait in Cerebral Palsy*, de Omar A. Galarraga C. et al. [Galarraga C, O. A., Vigneron, V., Dorizzi, B., Khouri, N., & Desailly, E. (2017). Predicting postoperative gait in cerebral palsy. *Gait & posture*, *52*, 45–51. <https://doi.org/10.1016/j.gaitpost.2016.11.012>], presenta un enfoque innovador para predecir la cinemática postoperatoria de las extremidades inferiores en pacientes con parálisis cerebral (PC). El estudio aprovecha la cinemática preoperatoria, los datos de la exploración física y la información quirúrgica para predecir los patrones cinemáticos tras la cirugía multinivel de evento único (SEMLS). Los autores emplean técnicas estadísticas de aprendizaje automático para desarrollar un sistema de predicción capaz de proporcionar estimaciones precisas de los ángulos de marcha postoperatorios para una diversa gama de procedimientos quirúrgicos y patrones de marcha. Por tanto, mientras que el primer artículo descrito en esta sección [O. A. Galarraga C., V. Vigneron, B. Dorizzi, N. Khouri & E. Desailly (2015) Predicting postoperative knee flexion during gait of cerebral palsy children, Computer Methods in Biomechanics and Biomedical Engineering, 18:sup1, 1940-1941, DOI: 10.1080/10255842.2015.1070583] se centra en la predicción de los ángulos de flexión de la rodilla, este nuevo estudio adopta un enfoque más exhaustivo para predecir los patrones generales de la marcha. No obstante, de manera similar a la investigación anteriormente mencionada [O. A. Galarraga C., V. Vigneron, B. Dorizzi, N. Khouri & E. Desailly (2015) Predicting postoperative knee flexion during gait of cerebral palsy children, Computer Methods in Biomechanics and Biomedical Engineering, 18:sup1, 1940-1941, DOI: 10.1080/10255842.2015.1070583], y con el objetivo de mejorar la precisión de la predicción y reducir la redundancia de datos, los autores utilizan el análisis de componentes principales (PCA) para reducir la dimensionalidad de las variables independientes, es decir, de los datos preoperatorios. A continuación, utilizan un único modelo de regresión lineal múltiple para predecir simultáneamente las múltiples variables cinemáticas postoperatorias. Es decir, este modelo de regresión tiene en cuenta una combinación de datos cinemáticos preoperatorios proyectados en un espacio de dimensión inferior mediante el análisis de componentes principales (PCA), datos del examen físico preoperatorio e información quirúrgica como características de entrada, y predice los resultados cinemáticos postoperatorios para una serie de ángulos cinemáticos. Finalmente, evalúan el rendimiento de la predicción mediante la raíz del error cuadrático medio (RMSE) para cada paciente del conjunto de prueba. Aunque la media postoperatoria resulta ser un predictor fiable para los ángulos del plano frontal, mostrando errores de predicción medios y desviaciones estándar mínimos, las predicciones para la dorsiflexión del tobillo no están bien estimadas y tienden a coincidir estrechamente con la cinemática media postoperatoria. Esto indica que el rendimiento del modelo para predecir la dorsiflexión del tobillo es menos preciso en comparación con otros parámetros cinemáticos, especialmente en comparación con los ángulos del plano sagital.

Los artículos analizados [Galarraga C, O. A., Vigneron, V., Dorizzi, B., Khouri, N., & Desailly, E. (2017). Predicting postoperative gait in cerebral palsy. *Gait & posture*, *52*, 45–51. <https://doi.org/10.1016/j.gaitpost.2016.11.012>], [Galarraga, Omar & Vigneron, Vincent & Dorizzi, Bernadette & Khouri, Nejib & Desailly, Eric. (2015). Estimation of Postoperative Knee Flexion at Initial Contact of Cerebral Palsy Children using Neural Networks. ICPRAM 2015 - 4th International Conference on Pattern Recognition Applications and Methods, Proceedings. 2. 10.5220/0005286503380342.], [Galarraga C, O. A., Vigneron, V., Dorizzi, B., Khouri, N., & Desailly, E. (2017). Predicting postoperative gait in cerebral palsy. *Gait & posture*, *52*, 45–51. <https://doi.org/10.1016/j.gaitpost.2016.11.012>] ponen de manifiesto que la aplicación del aprendizaje automático para predecir resultados relacionados con los patrones de la marcha es muy prometedora, y destacan el poder de estas técnicas para proporcionar información valiosa sobre la planificación del tratamiento y la toma de decisiones. No obstante, también destacan la presencia de algunos de los desafíos existentes: la estimación precisa de la dorsiflexión del tobillo, que es un parámetro crucial en el análisis de la marcha, sigue siendo un reto importante, lo que plantea áreas con potencial para una investigación adicional.

Sin embargo, si bien ninguna de las investigaciones anteriores se enfoca directamente en la predicción del progreso de la terapia, sientan un precedente en el uso del aprendizaje automático para predecir el progreso y la evolución de los pacientes con deficiencias musculoesqueléticas sometidos a rehabilitación.

# Marco teórico

La evaluación en el conjunto de prueba es esencial para determinar cómo los modelos generalizan a datos no vistos y su capacidad para hacer predicciones precisas en el mundo real.

Los indicadores clave como el la raíz del Error Cuadrático Medio (RMSE), el coeficiente de correlación (Pearson's r) y el coeficiente de determinación (R-squared) nos brindan una comprensión holística de la calidad de las predicciones.

# Métodos

## Herramientas y software

En el presente estudio se empleó el lenguaje de programación Python (versión 3.9.16) para el procesamiento de datos, el análisis y el desarrollo de modelos de aprendizaje automático. Se eligió Python por su amplia gama de bibliotecas y herramientas adaptadas a las aplicaciones de la ciencia de datos y el aprendizaje automático.

Se utilizaron las siguientes bibliotecas clave para facilitar diversos aspectos de la investigación:

* Pandas (versión 2.0.2): Pandas se empleó para manipular, manejar y preprocesar datos de forma eficiente [pandas].
* NumPy (versión 1.25.0): la biblioteca NumPy desempeñó un papel fundamental en los cálculos numéricos y las manipulaciones de matrices ya que permite realizar operaciones en matrices y *arrays* multidimensionales, esenciales para diversas tareas de análisis de datos [numpy].
* Matplotlib (versión 3.7.1) y Seaborn (versión 0.12.2): estas bibliotecas de visualización se utilizaron para crear distintos tipos de gráficos para el análisis exploratorio de datos y la presentación de resultados. Matplotlib proporciona un marco flexible para la creación de visualizaciones estáticas, interactivas y de calidad [matplotlib], mientras que Seaborn simplifica la creación de gráficos haciéndolos más estéticos [seaborn].
* SciPy (versión 1.11.1): la biblioteca SciPy se utilizó para realizar cálculos estadísticos [scipy].
* Statsmodels (versión 0.14.0): statsmodels se utilizó para la construcción de modelos, la evaluación de la heteroscedasticidad y la preparación de los datos para el análisis de regresión lineal [statsmodels].
* Scikit-Learn (versión 1.3.0): la biblioteca de aprendizaje automático Scikit-Learnofrece una amplia gama de algoritmos regresión, reducción dimensional, etc [scikit-learn]. Es por ello que se empleó para la selección, el entrenamiento y la evaluación de modelos, así como para la reducción de la dimensionalidad de los datos de entrada.
* XGBoost (versión 1.7.6): XGBoost es una biblioteca distribuida optimizada de *gradient boosting* diseñada para ser altamente eficiente, flexible y portable [xgboost]. Es por ello que se utilizó para construir y ajustar modelos de regresión en el marco del *Gradient Boosting.*

A parte de las bibliotecas anteriormente mencionadas, se adoptaron herramientas y prácticas específicas para garantizar un flujo de trabajo eficaz y organizado. Se eligió Visual Studio Code como principal entorno de desarrollo integrado para la codificación y gestión del proyecto, ya que ofrece una interfaz fácil de usar y mejora el proceso de desarrollo. Para aislar las dependencias del proyecto y garantizar la coherencia entre distintos entornos, se creó un entorno virtual utilizando el módulo *venv*. Esto ayudó a gestionar las versiones de las bibliotecas y paquetes necesarios para el proyecto, minimizando los posibles conflictos. Además, se empleó Git como sistema de control de versiones para realizar un seguimiento de los cambios, colaborar y gestionar las distintas versiones del proyecto [git]. GitHub, una plataforma de desarrollo colaborativo, se utilizó para alojar el repositorio del proyecto y facilitar el control de versiones [github].

## El conjunto de datos GaitRec

### Introducción al conjunto de datos GaitRec

El conjunto de datos GaitRec, presentado por Horsak et al. [gaitrec], es un *dataset* a gran escala, exhaustivo y ampliamente etiquetado, diseñado para facilitar la investigación en el análisis de las fuerzas de reacción del suelo (GRF) durante la locomoción humana. Este conjunto de datos proviene de una base de datos clínica de la marcha mantenida por un centro de rehabilitación austriaco. La recopilación de datos se llevó a cabo entre 2007 y 2018, con la participación de pacientes con diversas discapacidades musculoesqueléticas y controles sanos. Es importante destacar que, antes de publicar del conjunto de datos, se obtuvo la aprobación del Comité de Ética local de Baja Austria (GS1-EK-4/299-2014).

### Preparación y procesamiento

Antes de llevar a cabo las sesiones de análisis de la marcha, cada participante se sometió a un riguroso examen físico realizado por un médico.

Para la obtención de los datos, se solicitó a los participantes que caminaran sin ayuda a una velocidad de marcha elegida por ellos mismos en una pasarela de aproximadamente 10 metros que contaba con dos placas de fuerza incrustadas en su centro. Estas placas se coloraron en orden secuencial y al ras del suelo. Durante cada sesión, los sujetos caminaron hasta obtener un número mínimo de grabaciones válidas, que en la mayoría de los casos fue de diez grabaciones. Se consideraron grabaciones válidas cuando los pacientes caminaban de manera natural y había un claro impacto del pie en cada placa de fuerza. Los pacientes caminaron descalzos, con su calzado ortopédico o calzado normal, y en algunos casos con plantillas ortopédicas. En el caso de los controles sanos, caminaron descalzos o con su calzado normal.

Se capturaron tres señales analógicas de fuerza de reacción del suelo (componente vertical, anterior-posterior y medio-lateral), así como el centro de presión (COP), que posteriormente se convirtieron en señales digitales. El centro de presión y las fuerzas de reacción del suelo se registraron en el sistema de coordenadas de la placa de fuerza [gaitrec].

Para facilitar el análisis, se estandarizó la orientación de las señales medio-lateral y anteroposterior en todos los datos, de manera que las fuerzas medial y anterior siempre se representaran como valores positivos.

Debido a los protocolos internos del centro, los datos en bruto presentaban ruido al principio y al final de las señales. Con el propósito solucionar este problema y mejorar la calidad de los datos, se aplicó un umbral de 25 N a todos los datos de fuerza y luego se calculó el centro de presión. Estos datos son las señales de fuerza de reacción del suelo sin procesar (en bruto). Además, se generaron datos postprocesados *ready to use*. Para ello, el centro de presión solo se calculó cuando la fuerza vertical alcanzó los 80 N a fin de evitar imprecisiones en el cálculo del centro de presión en valores de fuerza pequeños [gaitrec].

A continuación, se centraron las coordenadas medio-laterales del centro de presión en el valor medio, mientras que las coordenadas anterior-posterior se centraron en cero. Las señales de fuerza procesadas se filtraron mediante un filtro de paso bajo de segundo orden con una frecuencia de corte de 20 Hz, a fin de reducir el ruido, y se normalizaron en el tiempo al 100% de la fase de apoyo (*stance* en inglés) del ciclo de la marcha, es decir, 101 puntos. Los valores de amplitud de las tres componentes de fuerza se expresaron como múltiplos del peso corporal (BW), dividiendo la fuerza por el producto de la masa corporal y la aceleración debida a la gravedad [gaitrec].

Tanto la normalización de la amplitud como la normalización temporal son operaciones necesarias para reducir los efectos de las covariables (como la antropometría) en las señales y reducir las diferencias temporales que dificultan las comparaciones de diferentes pasos [*Mullineaux, D. R., Milner, C. E., Davis, I. S. & Hamill, J. Normalization of ground reaction forces. J. Appl. Biomech. 22, 230–233 (2006).*], [*Mullineaux, D. R., Milner, C. E., Davis, I. S. & Hamill, J. Normalization of ground reaction forces. J. Appl. Biomech. 22, 230–233 (2006).*]*,* [*Helwig, N. E., Hong, S., Hsiao-Wecksler, E. T. & Polk, J. D. Methods to temporally align gait cycle data. J. Biomech. 44, 561–566 (2011).*].

Además, implementaron el algoritmo propuesto por Sangeux y Polak [*Sangeux, M. & Polak, J. A simple method to choose the most representative stride and detect outliers. Gait Posture 41, 726–730 (2015).*] para eliminar cualquier valor atípico antes de incluirlo en el conjunto de datos.

### Descripción y anotaciones

El conjunto de datos GaitRec contiene mediciones anonimizadas de fuerzas de reacción del suelo de un total de 2085 pacientes con deficiencias musculoesqueléticas (*gait disorders*, GD) y 211 controles sanos (HC).

Cada registro incluye metadatos adicionales como el identificador del sujeto, el identificador de la sesión, etiquetas que proporcionan información sobre la afectación específica, la edad, el sexo, el tipo de calzado, la velocidad de la marcha, si ese registro pertenece al conjunto de entrenamiento (balanceado o no) o al conjunto de prueba, etc. La descripción de todos los metadatos se muestra en la Tabla X.

Tabla X. Descripción de la información guardada en el fichero de metadatos [gaitrec].

|  |  |  |  |
| --- | --- | --- | --- |
| **Categorías/Variables** | **Formato** | **Unidad** | **Descripción** |
| **Identificadores** |  |  |  |
| SUBJECT\_ID | *integer* | — | Identificador único de un sujeto |
| SESSION\_ID | *integer* | — | Identificador único de una sesión |
| **Etiquetas** |  |  |  |
| CLASS\_LABEL | *string* | — | Etiquetas de clase anotadas |
| CLASS\_LABEL\_DETAILED | *string* | — | Etiquetas de clase anotadas para subclases |
| **Metadatos de los sujetos** |  |  |  |
| SEX | *binary* | — | femenino = 0  masculino = 1 |
| AGE | *integer* | años | Edad en la fecha de registro |
| HEIGHT | *integer* | centímetro | Altura corporal en centímetros |
| BODY\_WEIGHT | *double* | Kg·m  s2 | Peso corporal en Newton |
| BODY\_MASS | *double* | kg | Masa corporal |
| SHOE\_SIZE | *double* | EU | Número de calzado en el Sistema Europeo Continental |
| AFFECTED\_SIDE | *integer* | — | izquierda = 0  derecha = 1  ambas = 2 |
| **Metadata de los paseos o *trials*** |  |  |  |
| SHOD\_CONDITION | *integer* | — | descalzo y calcetines = 0 zapato normal = 1  zapato ortopédico = 2 |
| ORTHOPEDIC\_INSOLE | *binary* | — | sin plantilla = 0  con plantilla = 1 |
| SPEED | *integer* | — | Velocidad de la marcha:  lenta = 1  autoseleccionada = 2  rápida = 3 |
| READMISSION | *integer* | — | indica el número de reingresos = 0 ... n |
| SESSION\_TYPE | *integer* | — | medición inicial = 1 medición de control = 2 medición inicial tras readmisión = 3 |
| SESSION\_DATE | *string* | — | fecha de la sesión de registro en el formato "DD-MM-YYYY " |
| **Información de Train-Test Split** |  |  |  |
| TRAIN | *binary* | — | forma parte (=1) o no forma parte (=0) de TRAIN |
| TRAIN\_BALANCED | *binary* | — | forma parte (=1) o no forma parte (=0) de TRAIN\_BALANCED |
| TEST | *binary* | — | forma parte (=1) o no forma parte (=0) de TEST |

Los individuos clasificados como controles sanos se seleccionaron bajo la condición de no experimentar dolor o molestias en las extremidades inferiores ni en la columna vertebral, además de no llevar ningún tipo de ortesis ni plantillas ortopédicas. Se excluyeron aquellos que tenían antecedentes de cirugía o traumatismo en la columna vertebral o las extremidades inferiores [gaitrec].

La duración habitual de la estancia de los pacientes en el centro de rehabilitación varía desde algunos días hasta varias semanas, influenciada por elementos como el diagnóstico, el tratamiento, la terapia, la cirugía realizada y el avance en la recuperación. Durante este período de tiempo, los pacientes se someten periódicamente a análisis de la marcha, normalmente una vez por semana. Cuando los pacientes ingresan en el centro de rehabilitación, se establecen los resultados esperados de la terapia en colaboración con el terapeuta. Una vez que los objetivos se logran total o parcialmente, se procede a dar de alta a los pacientes, dándoles la opción de ingresar de nuevo si fuese necesario.

El presente conjunto de datos contiene los datos recopilados durante toda la estancia o estancias de cada paciente, abarcando así todo el progreso de rehabilitación de los pacientes [gaitrec]. El *dataset* fue etiquetado manualmente por un fisioterapeuta experimentado basándose en el diagnóstico médico disponible de cada paciente, y se clasifica jerárquicamente en una clase de controles sanos (HC) y cuatro clases basadas en el nivel de afectación de la articulación anatómica: cadera (H), rodilla (K), tobillo (A) y calcáneo (C). Las etiquetas de anotación proporcionan información sobre la afectación específica y su localización anatómica. En la Tabla X se muestra el resumen demográfico del conjunto de datos y las clases predefinidas. Esta extensa anotación permite un análisis en profundidad y la clasificación de los patrones de la marcha a través de diferentes tipos de deficiencias.

Tabla X. Resumen demográfico del conjunto de datos y las clases predefinidas [gaitrec].

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Clases** | **N** | **Edad (años) Media (SD)** | **Masa corporal (kg)**  **Media (SD)** | **Género**  **(m/f)** | **Número de paseos o *trials* bilaterales** |
| Control sano | 211 | 34.7 (13.9) | 73.9 (15.6) | 104/107 | 7755 |
| Cadera | 450 | 42.6 (12.8) | 73.9 (15.6) | 373/77 | 12748 |
| Rodilla | 625 | 41.6 (12.0) | 84.3 (18.6) | 426/199 | 19873 |
| Tobillo | 627 | 41.6 (11.4) | 87.0 (18.0) | 498/129 | 21386 |
| Calcáneo | 382 | 43.5 (10.4) | 84.0 (14.5) | 339/43 | 13970 |
| **Total** | **2295** | **41.5 (12.1)** | **83.6 (17.3)** | **1740/555** | **75732** |

### Estructura de los datos

Todos datos se encuentran disponibles en formato de valores separados por comas (CSV). Los veinte archivos de datos de fuerzas de reacción del suelo (GRF) están organizados siguiendo una convención de nomenclatura "GRF-tipo-procesamiento-lado.csv" como se muestra en la Tabla X\_(Descripción de los datos almacenados en los ficheros). Aquí, "tipo" indica si el archivo contiene series temporales verticales ("F\_V"), anteroposteriores ("F\_AP"), medio laterales ("F\_ML"), o centro de presión (COP) anteroposteriores o medio laterales ("COP\_AP", "COP\_ML"). El término "procesamiento" señala si los archivos contienen datos en bruto ("RAW") o datos postprocesados ("PRO"). Finalmente, "lado" indica si los datos provienen del lado derecho o izquierdo del cuerpo. Por tanto, se proporcionan los datos en dos formatos. Por un lado, datos en formato original (datos brutos o crudos), y por otro lado, datos postprocesados sometidos a procesos de reducción del ruido y normalización. Esto da flexibilidad a futuros investigadores, permitiéndoles utilizar los datos crudos y aplicar sus propios procesamientos, o emplear los datos postprocesados directamente.

Los archivos "GRF-tipo-procesamiento-lado.csv" se estructuran de manera matricial con N filas × M columnas. En cada fila se encuentran los datos correspondientes a un sujeto y una sesión de grabación específica. La primera columna identifica al sujeto ("SUBJECT\_ID"), la segunda a la sesión de grabación ("SESSION\_ID"), y la tercera al paseo o *trial* individual dentro de dicha sesión ("TRIAL\_ID"). Del mismo modo, el archivo de metadatos ("GRF-metadata.csv") que contiene anotaciones y datos adicionales relacionados con el tema, se estructura en una matriz con N filas × M columnas (véase la Tabla X).

Con el propósito de poder evaluar con precisión el rendimiento de los modelos de aprendizaje automático, en el *dataset* vine incorporada una división predefinida y aleatoria en tres subconjuntos para su uso en el entrenamiento y evaluación. El conjunto de datos GaitRec se segmenta en un conjunto de entrenamiento desequilibrado (TRAIN) y un conjunto de prueba (TEST). El primero se emplea para entrenar y perfeccionar modelos de aprendizaje automático utilizando métodos como la validación cruzada, mientras que el segundo se destina a la evaluación definitiva. No obstante, la disparidad entre las clases en el conjunto de entrenamiento puede influir negativamente en la optimización de los modelos. Para abordar esto, los autores crearon un subconjunto equilibrado de entramiento, denominado TRAIN\_BALANCED, que incluye exclusivamente los datos correspondientes a las evaluaciones iniciales (primera sesión de medición), en la cual se contabilizan al menos cinco paseos o *trials* para cada lado del cuerpo por sesión.

Tabla X. Descripción de los datos almacenados en los ficheros "GRF\_\*.csv". *n* es el número de fotogramas durante un paso a través de la placa de fuerza para los datos sin procesar ("RAW") o un vector normalizado en el tiempo de 101 puntos para los datos postprocesados ("PRO") [gaitrec].

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Variables** | **Fichero asociado** | **Formato** | **Dimensión** | **Unidad** | **Descripción** |
| Vertical GRF | GRF\_F\_V-RAW\_\*.csv | *double* | 1 × n | Newton | Fuerza vertical de reacción al suelo bruta |
| Anterior-posterior GRF | GRF\_F\_AP-RAW\_\*.csv | *double* | 1 × n | Newton | Fuerza de cizallamiento de frenado y propulsión bruta |
| Medio-lateral GRF | GRF\_F\_ML\_RAW\_\*.csv | *double* | 1 × n | Newton | Fuerza de cizallamiento medio-lateral bruta |
| COP anterior-posterior | GRF\_COP\_AP\_RAW\_\*.csv | *double* | 1 × n | Centímetro | Coordenada COP bruta en dirección de marcha |
| COP medio-lateral | GRF\_COP\_ML\_RAW\_\*.csv | *double* | 1 × n | Centímetro | Coordenada COP bruta en dirección mediolateral |
| Vertical GRF | GRF\_F\_V\_PRO\_\*.csv | *double* | 1 × n | Múltiplo del peso corporal | Fuerza de reacción vertical al suelo postprocesada |
| Anterior-posterior GRF | GRF\_F\_AP\_PRO\_\*.csv | *double* | 1 × n | Múltiplo del peso corporal | Fuerza de cizallamiento de frenado y propulsión postprocesada |
| Medio-lateral GRF | GRF\_F\_ML\_PRO\_\*.csv | *double* | 1 × n | Múltiplo del peso corporal | Fuerza de cizallamiento medio-lateral postprocesada |
| COP anterior-posterior | GRF\_COP\_AP\_PRO\_\*.csv | *double* | 1 × n | % apoyo (*stance)* | Coordenada COP postprocesada en la dirección de la marcha |
| COP medio-lateral | GRF\_COP\_ML\_PRO\_\*.csv | *double* | 1 × n | % apoyo (*stance)* | Coordenada COP post-procesada en dirección medio-lateral |

## Construcción de los conjuntos de datos

Para poder realizar nuestra investigación, primero de todo tuvimos que crear, a partir de los diferentes archivos en formato CSV que proporciona GaitRec, nuestros conjuntos de datos personalizados para que se adaptase a los objetivos de nuestro proyecto.

Decidimos trabajar con el conjunto de datos postprocesados (“PRO”) sometidos a procedimientos de eliminación de ruido y normalización. Trabajar con el conjunto de datos postprocesado, más limpio y fiable, coherente y estandarizado, y sin valores atípicos ofrece una serie de ventajas. Algunas de ellas son el ahorro de recursos computacionales y tiempo, la facilidad de reproducir nuestro estudio, así como comparar y contrastar los resultados obtenidos con otras futuras investigaciones que utilicen los mismos métodos de preprocesamiento o similares.

Optamos por trabajar con el conjunto de entrenamiento desbalanceado (“TRAIN”) en lugar de utilizar el subconjunto balanceado ("TRAIN\_BALANCED"). El subconjunto balanceado (“TRAIN\_BALANCED”) comprende sólo los datos de las evaluaciones iniciales (primera sesión de medición), que contienen al menos cinco paseos o *trials* para cada lado del cuerpo por sesión [gaitrec]. Nuestra elección se basa en la necesidad de considerar la variabilidad y la evolución completa de los datos a lo largo del proceso de terapia y rehabilitación.

A continuación, creamos un archivo CSV con los metadatos de los pacientes pertenecientes solamente al conjunto de entrenamiento no balanceado, y otro archivo CSV con los metadatos de los pacientes pertenecientes solamente al conjunto de prueba.

Los archivos "GRF\_tipo\_procesamiento\_lado.csv", que contienen los datos postprocesados para cada una de las extremidades inferiores de los pacientes presentes en el conjunto de entrenamiento y prueba, se guardan en la carpeta “postprocessed” (véase la Figura X\_postprocessed).

A screenshot of a computer

Description automatically generated

Figura X\_postprocessed. Contenido de la carpeta “postprocessed”.

Con el objetivo de separar los pacientes pertenecientes al conjunto de entrenamiento de los pertenecientes al conjunto de prueba, creamos las carpetas “posprocesed\_train” y “test\_set” a partir de la carpeta “postprocessed”. En estas carpetas guardamos los archivos "GRF\_tipo\_procesamiento\_lado.csv" correspondientes a cada uno de los conjuntos de entrenamiento y prueba, respectivamente (véase la Figura X\_postptrain\_testset).

A screenshot of a computer program

Description automatically generated A screenshot of a computer

Description automatically generated

Figura X\_postptrain\_testset. Contenido de las carpetas “postprocesed\_train” y “test\_set”.

Con el fin de alcanzar el objetivo propuesto y poder realizar predicciones del progreso de la rehabilitación, se estableció un requisito mínimo de dos sesiones por paciente. En consecuencia, se excluyó a aquellos pacientes que habían participado en solo una sesión de terapia.

Cabe destacar también, la importancia de crear un subconjunto de individuos que fuese más o menos homogéneo, ya que esto contribuye a minimizar la variabilidad inherente en la muestra, lo que aumenta la fiabilidad y precisión de los análisis realizados. Además, al focalizar en grupos más homogéneos, se facilita la identificación de patrones y tendencias que podrían quedar disueltas en conjuntos más heterogéneos. Para ello, clasificamos los sujetos en función del número de sesiones y de la clase. Para tener los sujetos clasificados, filtramos en función del número deseado de sesiones y clase, y guardamos los datos filtrados en carpetas específicas de grupos y sesiones para su posterior procesamiento o análisis. Inicialmente, seleccionamos todos los grupos (tobillo, calcáneo, cadera, controles sanos, y rodilla), así como los sujetos que habían realizado 2, 3 y 4 sesiones de terapia, ya que estos números de sesiones de rehabilitación eran los más comunes entre la mayoría de los sujetos. La estructura de las carpetas se muestra en la Figura X\_groups\_sessions\_data\_train\_test y en la Tabla X\_groups\_sessions\_data\_train\_test.

Finalmente, se tomó la decisión de comenzar por un enfoque específico, eligiendo el grupo "tobillo" (A) como punto de partida y empleando dos sesiones de terapia (véase Tabla X\_A\_group\_2sessions\_data\_train\_test). Esta elección se basó en la necesidad de iniciar el análisis exploratorio de manera focalizada y controlada. Iniciar con un grupo y un número de sesiones determinados permitió una exploración detallada de los patrones y características presentes en esos datos, así como una evaluación de los procedimientos de análisis. Además, esta selección proporcionó una oportunidad para familiarizarse con los datos y comprender mejor sus particularidades, contribuyendo a la toma de decisiones informadas en la posterior adaptación de los métodos y enfoques. Esta elección permitió sentar un punto de partida práctico y manejable en el proceso investigativo.

A screenshot of a computer program

Description automatically generatedA screenshot of a computer program

Description automatically generated

Tabla X\_A\_group\_2sessions\_data\_train\_test.

|  |  |  |  |
| --- | --- | --- | --- |
| **Datos** | **N.º de sesiones** | **Grupo o clase** | **Ficheros** |
| TRAIN | 2 | Tobillo (A) | GRF\_COP\_AP\_PRO\_left\_train\_A.csv  GRF\_COP\_AP\_PRO\_right\_train\_A.csv  GRF\_COP\_ML\_PRO\_left\_train\_A.csv  GRF\_COP\_ML\_PRO\_right\_train\_A.csv  GRF\_F\_AP\_PRO\_left\_train\_A.csv  GRF\_F\_AP\_PRO\_right\_train\_A.csv  GRF\_F\_ML\_PRO\_left\_train\_A.csv  GRF\_F\_ML\_PRO\_right\_train\_A.csv  GRF\_F\_V\_PRO\_left\_train\_A.csv  GRF\_F\_V\_PRO\_right\_train\_A.csv |
| TEST | 2 | Tobillo (A) | GRF\_COP\_AP\_PRO\_left\_test\_A.csv  GRF\_COP\_AP\_PRO\_right\_test\_A.csv  GRF\_COP\_ML\_PRO\_left\_test\_A.csv  GRF\_COP\_ML\_PRO\_right\_test\_A.csv  GRF\_F\_AP\_PRO\_left\_test\_A.csv  GRF\_F\_AP\_PRO\_right\_test\_A.csv  GRF\_F\_ML\_PRO\_left\_test\_A.csv  GRF\_F\_ML\_PRO\_right\_test\_A.csv  GRF\_F\_V\_PRO\_left\_test\_A.csv  GRF\_F\_V\_PRO\_right\_test\_A.csv |

Tabla X\_groups\_sessions\_data\_train\_test.

|  |  |  |  |
| --- | --- | --- | --- |
| **Conjunto de datos** | **Número de sesiones** | **Grupo o clase** | **Ficheros** |
| TRAIN | 2 | Tobillo  (A) | GRF\_COP\_AP\_PRO\_left\_train\_A.csv  GRF\_COP\_AP\_PRO\_right\_train\_A.csv  GRF\_COP\_ML\_PRO\_left\_train\_A.csv  GRF\_COP\_ML\_PRO\_right\_train\_A.csv  GRF\_F\_AP\_PRO\_left\_train\_A.csv  GRF\_F\_AP\_PRO\_right\_train\_A.csv  GRF\_F\_ML\_PRO\_left\_train\_A.csv  GRF\_F\_ML\_PRO\_right\_train\_A.csv  GRF\_F\_V\_PRO\_left\_train\_A.csv  GRF\_F\_V\_PRO\_right\_train\_A.csv |
| Calcáneo  (C) | GRF\_COP\_AP\_PRO\_left\_train\_C.csv  GRF\_COP\_AP\_PRO\_right\_train\_C.csv  GRF\_COP\_ML\_PRO\_left\_train\_C.csv  GRF\_COP\_ML\_PRO\_right\_train\_C.csv  GRF\_F\_AP\_PRO\_left\_train\_C.csv  GRF\_F\_AP\_PRO\_right\_train\_C.csv  GRF\_F\_ML\_PRO\_left\_train\_C.csv  GRF\_F\_ML\_PRO\_right\_train\_C.csv  GRF\_F\_V\_PRO\_left\_train\_C.csv  GRF\_F\_V\_PRO\_right\_train\_C.csv |
| Cadera  (H) | GRF\_COP\_AP\_PRO\_left\_train\_H.csv  GRF\_COP\_AP\_PRO\_right\_train\_H.csv  GRF\_COP\_ML\_PRO\_left\_train\_H.csv  GRF\_COP\_ML\_PRO\_right\_train\_H.csv  GRF\_F\_AP\_PRO\_left\_train\_H.csv  GRF\_F\_AP\_PRO\_right\_train\_H.csv  GRF\_F\_ML\_PRO\_left\_train\_H.csv  GRF\_F\_ML\_PRO\_right\_train\_H.csv  GRF\_F\_V\_PRO\_left\_train\_H.csv  GRF\_F\_V\_PRO\_right\_train\_H.csv |
| Control Sano  (HC) | GRF\_COP\_AP\_PRO\_left\_train\_HC.csv  GRF\_COP\_AP\_PRO\_right\_train\_HC.csv  GRF\_COP\_ML\_PRO\_left\_train\_HC.csv  GRF\_COP\_ML\_PRO\_right\_train\_HC.csv  GRF\_F\_AP\_PRO\_left\_train\_HC.csv  GRF\_F\_AP\_PRO\_right\_train\_HC.csv  GRF\_F\_ML\_PRO\_left\_train\_HC.csv  GRF\_F\_ML\_PRO\_right\_train\_HC.csv  GRF\_F\_V\_PRO\_left\_train\_HC.csv  GRF\_F\_V\_PRO\_right\_train\_HC.csv |
| Rodilla  (K) | GRF\_COP\_AP\_PRO\_left\_train\_K.csv  GRF\_COP\_AP\_PRO\_right\_train\_K.csv  GRF\_COP\_ML\_PRO\_left\_train\_K.csv  GRF\_COP\_ML\_PRO\_right\_train\_K.csv  GRF\_F\_AP\_PRO\_left\_train\_K.csv  GRF\_F\_AP\_PRO\_right\_train\_K.csv  GRF\_F\_ML\_PRO\_left\_train\_K.csv  GRF\_F\_ML\_PRO\_right\_train\_K.csv  GRF\_F\_V\_PRO\_left\_train\_K.csv  GRF\_F\_V\_PRO\_right\_train\_K.csv |
| 3 | A |  |
| C |  |
| H |  |
| HC |  |
| K |  |
| 4 | A |  |
| C |  |
| H |  |
| HC |  |
| K |  |
| TEST | 2 | A |  |
| C |  |
| H |  |
| HC |  |
| K |  |
| 3 | A |  |
| C |  |
| H |  |
| HC |  |
| K |  |
| 4 | A |  |
| C |  |
| H |  |
| HC |  |
| K |  |

El siguiente paso consistió en eliminar las sesiones intermedias y retener únicamente la primera y la última sesión de cada paciente. Para cada paciente y sesión, se calculó el valor promedio de los diferentes *trials* realizados en esa sesión. Esto permitió obtener una medida representativa de cada sesión para su posterior análisis.

Para realizar una mayor acotación en nuestro estudio, focalizamos nuestra atención en una variable concreta. En este sentido, decidimos centrarnos en la fuerza de reacción del suelo en dirección anterior-posterior (fuerza de cizallamiento de frenado y propulsión), es decir, los archivos "GRF\_F\_AP\_PRO\_right\_train\_A.csv" y "GRF\_F\_AP\_PRO\_left\_train\_A.csv". Una de las razones clave detrás de esta elección fue la naturaleza de las líneas promedio de la primera y última sesión de rehabilitación, que representan la evolución del promedio del peso corporal durante la fase de apoyo (porcentaje de la fase de apoyo o *stance*) en las sesiones de terapia. Estas líneas claramente exhiben un punto mínimo y uno máximo distintivos. Al centrarnos en estos valores extremos, logramos gráficos más comprensibles y con resultados visiblemente notorios. Elegimos esta variable con la intención de trabajar con datos que fueran más fáciles de interpretar y con un impacto visual más alto, lo que a su vez mejoró nuestra capacidad para analizar y extraer información valiosa.

Dado que disponíamos de dos archivos, uno correspondiente a cada extremidad inferior ("GRF\_F\_AP\_PRO\_right\_train\_A.csv" y "GRF\_F\_AP\_PRO\_left\_train\_A.csv"), fue necesario combinarlos. De esta manera unificamos los datos de ambas extremidades inferiores en un único conjunto de datos. Abordamos esta integración considerando cada pierna como una observación independiente. Así que optamos por unir el conjunto de datos de la pierna izquierda debajo del conjunto de datos de la pierna derecha (R), creando un patrón “RRRLLL”. Esta aproximación nos permitió tratar cada extremidad inferior como si fuera un paciente diferente, ampliando así la cantidad de datos disponibles.

El resultado fueron dos conjuntos de datos separados, uno que contenía los datos de entrenamiento y otro los datos de prueba (véase la Figura X\_train\_combined). Estos conjuntos de datos contienen el identificador único del sujeto (“SUBJECT\_ID”), el identificador único de una sesión (“SESSION\_ID”), y la media de los identificadores de los *trials* o paseos realizados en esa sesión ("TRIAL\_ID"). Estas tres primeras columnas inicialmente se mantuvieron en la estructura inicial. Luego, se eliminaron debido a que contribuían al proceso de predicción. Las siguientes columnas (de “F\_AP\_PRO\_1” a “F\_AP\_PRO\_101”) representan el valor promedio de los diferentes *trials* realizados en cada sesión por cada paciente para cada uno de los miembros inferiores (MMII). Es decir, representan los valores promedio de la fuerza de reacción del suelo en dirección anterior-posterior (fuerza de cizallamiento de frenado y propulsión) en diferentes puntos durante la fase de apoyo del ciclo de la marcha, para todos los *trials* de esa sesión en particular para cada sujeto. Cada fila del *dataset* representa una observación de un sujeto en una sesión particular. La primera fila correspondiente a un paciente representa su primera sesión, mientras que la segunda fila representa la última sesión de ese mismo paciente. Además, los datos de cada fila están divididos en dos partes: la primera parte se refiere al miembro inferior derecho, mientras que la segunda parte se refiere al miembro inferior izquierdo. Es importante recordar que tomamos la decisión de unir los datos de la pierna izquierda (L) debajo de los datos de la pierna derecha (R), lo que resulta en un patrón "RRRLLL". Esto significa que primero se presentan los datos del miembro inferior derecho (R) y luego los datos del miembro inferior izquierdo (L).

A table of numbers with numbers

Description automatically generated with medium confidence

Figura X\_train\_combined.

Cabe destacar que todos los pasos previamente descritos para construir los conjuntos de datos personalizados, de manera que se ajustaran a los objetivos específicos de nuestro proyecto, fueron unificados y encapsulados en funciones. Esta aproximación permitió la aplicación sistemática de dichos pasos tanto a los datos de entrenamiento como a los datos de prueba, otorgando una mayor eficiencia al proceso en su conjunto. El pipeline completo que abarca la construcción de los conjuntos de datos puede visualizarse de manera esquemática en la Figura X\_construccion\_dataset\_pipeline.

Falta crear el esquema del pipeline de construcción de los conjuntos de datos

## Análisis y comprobaciones iniciales

Mientras construíamos nuestro conjunto de datos, realizamos un análisis inicial previo y una serie de comprobaciones.

En el archivo CSV que contiene los metadatos de los pacientes pertenecientes solamente al conjunto de entrenamiento no balanceado, consultamos el número de pacientes por grupo o clase y el número de pacientes únicos para cada subclase.

En los archivos "GRF\_tipo\_procesamiento\_lado.csv" de la carpeta “postprocessed\_train”, correspondientes al conjunto de entrenamiento”, comprobamos si había pacientes que sólo tuvieran una sesión, o, por el contrario, que cada individuo tuviera varias sesiones. También consultamos el número máximo y mínimo de sesiones por paciente y visualizamos la distribución de los sujetos por número de sesiones. Por lo que respecta a los paseos o *trials*, comprobamos si había varios *trials* por paciente, consultamos el número máximo y mínimo de *trials* por paciente, y comprobamos que el número de *trials* para cada miembro inferior (derecho e izquierdo) fuese igual para cada paciente. A continuación, calculamos el número de *trials* por paciente y visualizamos los 50 sujetos con más *trials* y los 50 sujetos con menos *trials.* También visualizamos la distribución de los sujetos por número de *trials.*

Finalmente, visualizamos y comparamos los promedios de la primera y última sesión de pacientes con deficiencias musculoesqueléticas en el tobillo sometidos a 2 sesiones de rehabilitación en relación con los promedios de los controles sanos (HC). Para ello se utilizaron gráficos de líneas que representan la evolución de los promedios de peso corporal durante la fase de apoyo (% de la fase de apoyo o *stance*) en las sesiones de terapia. Esta visualización ayudó a identificar tendencias y patrones significativos en la evolución del peso corporal durante la fase de apoyo en diferentes sesiones de terapia.

## Estudio de la población y cohorte de pacientes

El conjunto de datos de entrenamiento creado para este estudio comprende un total de 90 pacientes únicos con deficiencias musculoesqueléticas en el tobillo, los cuales realizaron dos sesiones de terapia o rehabilitación. No obstante, consideramos cada miembro inferior como una observación independiente, lo que resulta en un total de 180 observaciones (N=180). De entre los 90 pacientes, 73 son hombres y 17 mujeres, lo que refleja una proporción de sexo desigual. La edad media de la cohorte es de 42.82 años, con una masa corporal promedio de 89.96 kg. Para más detalles sobre las características demográficas de la cohorte de pacientes, consultar la Tabla X\_demographic\_trainset.

Adicionalmente, el estudio incluye 102 controles sanos. Como en el caso de los pacientes, tratamos cada miembro inferior como una observación independiente, lo que resulta en un total de 204 observaciones (N=204). De estos 102 controles sanos, 45 son hombres y 57 mujeres, siendo esta una proporción de sexo más equilibrada en comparación con la cohorte de pacientes. La edad media de esta segunda cohorte es de 36.28 años, con una masa corporal promedio de 73.95 kg. Para más detalles sobre las características demográficas de la cohorte de controles sanos, consultar la Tabla X\_demographic\_HC.

Tabla X\_demographic\_trainset. Resumen demográfico del conjunto de datos de entrenamiento de pacientes con deficiencias musculoesqueléticas en el tobillo.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Dataset** | **Clase** | **N.º sesiones** | **N.º pacientes** | **N** | **Edad media**  **(años)** | **Masa corporal media (kg)** | **Género**  **(m/f)** |
| TRAIN | Tobillo (A) | 2 | 90 | 180 | 42.82 | 89.96 | 73/17 |

Tabla X\_demographic\_HC. Resumen demográfico del conjunto de datos los controles sanos.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Dataset** | **Clase** | **N.º sujetos** | **N** | **Edad media**  **(años)** | **Masa corporal media (kg)** | **Género**  **(m/f)** |
| TRAIN | Controles sanos (HC) | 102 | 204 | 36.28 | 73.95 | 45/57 |

## Análisis exploratorio de los datos

Para comenzar el análisis exploratorio de datos, llevamos a cabo una exploración y una serie de pasos iniciales para entender la naturaleza y calidad de nuestro conjunto de datos de entrenamiento.

En primer lugar, realizamos una exploración visual del conjunto de datos para tener una visión general de su estructura. Posteriormente, comprobamos la presencia de valores nulos o faltantes, asegurándonos de que los datos estuvieran completos para un análisis adecuado. También, calculamos estadísticas descriptivas para las variables numéricas, lo que nos permitió obtener información sobre tendencias centrales y dispersión en los datos. Además, verificamos la existencia de duplicados en los registros para asegurarnos de la integridad del *dataset*.

A continuación, nos enfocamos en comprender mejor la composición de las cohortes de pacientes y controles sanos. Realizamos un análisis específico para calcular el número total de pacientes y controles sanos incluidos en nuestro estudio. Al considerar cada miembro inferior como una observación independiente, determinamos el número total de observaciones en ambas cohortes. Calculamos la edad media en años, la masa corporal promedio en kilogramos, y examinamos la distribución de sexo en ambas cohortes para obtener una idea más precisa de las características demográficas de los sujetos en cada cohorte.

Para obtener una comprensión más profunda de la estructura y relaciones presentes en el conjunto de datos realizamos los siguientes pasos. Visualizamos la distribución de cada una de las 101 variables numéricas mediante *boxplots* e histogramas para analizar la dispersión de los datos y detectar posibles valores atípicos o patrones de distribución inusuales. Posteriormente, evaluamos la normalidad de los datos utilizando la prueba de Anderson-Darling. Esta prueba nos permitió determinar si las variables seguían o no una distribución normal. Además, realizamos un análisis de correlación entre todas las variables utilizando el coeficiente de correlación de Pearson. Aunque generalmente se recomienda escalar las variables antes de llevar a cabo un análisis de correlación, en nuestro caso decidimos no hacerlo ya que al trabajar con los datos postprocesados proporcionados por GaitRec, ya habían sido normalizados previamente.

Todo este proceso fue encapsulado en funciones para garantizar la modularidad y reutilización en nuestro análisis.

En resumen, el análisis exploratorio nos proporcionó una comprensión sólida de la estructura, calidad, distribución y relaciones en nuestro conjunto de datos, lo que sentó las bases para aplicar análisis más avanzados, técnicas de procesamiento de datos y poder realizar los modelos de aprendizaje automático.

## División de los datos

En pasos previos, realizamos la separación de los datos de entrenamiento y prueba en conjuntos de datos distintos. Sin embargo, antes de introducir los datos en cualquier modelo de aprendizaje automático, se deben separar las características y las variables objetivo o de salida. Esto es necesario para garantizar la integridad de los modelos de aprendizaje automático, prevenir la fuga de datos (*data leakage*) y poder hacer una evaluación precisa del rendimiento del modelo en datos no vistos. En nuestro caso, las variables predictoras son los valores de la primera sesión de terapia, localizados en las filas impares, mientras que las variables objetivo son los valores de la última sesión, ubicados en las filas pares.

El proceso de separación de variables predictoras y variables objetivo o de salida lo ejecutamos mediante la implementación de una función específica para dicha tarea. Esta función excluye las columnas “SUBJECT\_ID”, “SESSION\_ID” y “TRIAL\_ID”, y crea los *arrays* “X\_train”, “y\_train”, “X\_test” y “y\_test”. Para llevarlo a cabo, extrae las filas impares del conjunto de entrenamiento y las asigna a “X\_train”. A continuación, extrae las filas pares del conjunto de entrenamiento y las asigna a “y\_train”. Posteriormente, realiza el mismo procedimiento para los datos del conjunto de prueba.

## Preprocesamiento de los datos

Una vez tuvimos los datos en el formato adecuado para poder introducirlos en los modelos de aprendizaje automático, iniciamos la fase de preprocesamiento. En la etapa de preprocesamiento de los datos, implementamos varias aproximaciones con el objetivo de optimizar el rendimiento de los modelos de aprendizaje automático. Estas aproximaciones se aplicaron a los conjuntos de datos de entrada y salida, con el propósito de explorar cómo distintos preprocesamientos podrían impactar en las predicciones de los modelos.

### Primera aproximación: datos sin procesar

En la primera aproximación no se aplicó ningún tipo de procesamiento adicional a los datos de entrada y salida. Las predicciones se realizaron utilizando los datos en su forma original, sin modificar la dimensionalidad ni el contenido de las variables predictoras y/o objetivo.

### Segunda Aproximación: PCA en los datos de entrada (X)

La segunda aproximación consistió en aplicar el análisis de componentes principales (PCA) solo a los datos de entrada (X), que son los valores de la primera sesión de terapia. La idea detrás de esto es reducir la dimensionalidad de las variables predictoras para mejorar la precisión de las predicciones y eliminar redundancias. Aunque generalmente se recomienda escalar las variables antes de aplicar el PCA, en este caso no lo hicimos debido a que los datos postprocesados proporcionados por GaitRec ya habían sido normalizados previamente. Establecimos seleccionar los componentes principales que explicaran o retuvieran al menos el 90% de la varianza de los datos originales. La aplicación de esta segunda aproximación vino inspirada por el trabajo realizado por Omar A. Galarraga C. et al. descrito artículo científico titulado *Predicting Postoperative Gait in Cerebral Palsy* [Galarraga C, O. A., Vigneron, V., Dorizzi, B., Khouri, N., & Desailly, E. (2017). Predicting postoperative gait in cerebral palsy. *Gait & posture*, *52*, 45–51. <https://doi.org/10.1016/j.gaitpost.2016.11.012>] donde tratando de resolver un problema similar al nuestro, utilizaron el análisis de componentes principales para reducir la dimensionalidad de las variables independientes o de entrada con el objetivo de mejorar la precisión de la predicción y reducir la redundancia de datos.

### Tercera Aproximación: PCA en los datos de salida (Y)

En la tercera aproximación, aplicamos el PCA solo a los datos de salida (Y), correspondientes a los valores de la última sesión de terapia. De la misma forma que en la segunda aproximación, optamos por no escalar las variables antes de aplicar el PCA debido a la previa normalización de los datos, y establecimos nuevamente el objetivo de seleccionar los componentes principales que explicaran o retuvieran al menos el 90% de la varianza de los datos originales. Después de obtener las predicciones de los modelos de aprendizaje automático, transformamos esas predicciones de nuevo al espacio original de las variables objetivo utilizando el método *inverse\_transform()* proporcionado por la librería Scikit-Learn*.* Para volver al espacio original, en lugar de utilizar los componentes principales originales, utilizamos los componentes principales predichos. Es decir, teníamos los datos de salida del conjunto de prueba original (“y\_test”), hicimos un PCA, y nos fuimos a un espacio nuevo. Por tanto, podíamos volver al espacio original porque disponíamos de las combinaciones lineales. Con esas combinaciones lineales tuvimos unos componentes principales que posteriormente predecimos de nuevo.

### Cuarta aproximación: PCA en los datos de entrada y salida (X y Y)

En la cuarta y última aproximación, aplicamos el PCA tanto a los datos de entrada (X) como a los de salida (Y) con el objetivo de reducir la dimensionalidad de los datos de la primera y última sesión de terapia. Establecimos nuevamente el objetivo de seleccionar los componentes principales que explicaran o retuvieran al menos el 90% de la varianza de los datos originales. De la misma forma que en la tercera aproximación, después de obtener las predicciones con los modelos de aprendizaje automático, revertimos la transformación del PCA transformando los datos de nuevo al espacio original de las variables objetivo utilizando el método *inverse\_transform()* proporcionado por la librería Scikit-Learn

Cada uno de los algoritmos de aprendizaje automático utilizados en este estudio pasó por estas cuatro aproximaciones o tipos de preprocesamiento. En otras palabras, generamos cuatro modelos distintos para cada algoritmo, considerando cada uno de los tipos de preprocesamiento mencionados anteriormente. Esto nos permitió evaluar cómo cada aproximación afectaba al rendimiento de los modelos y nos permitió determinar cuál de ellas proporcionó las mejores predicciones. El proceso de generación y evaluación de estos modelos de aprendizaje automático se describe detalladamente en la siguiente sección.

## Análisis multivariante

Ideas:

* Intro
* Elección de algoritmos
* Pipeline general/común
  + Entrenamiento
    - Entrenar y validar usand CV
    - Hiperparametros
  + Evaluación
    - Explicar función común de evaluación
    - Funciones auxiliares para gráficar resultados
    - Análisis de residuos
* Algoritmos y modelos: explicar particularidades de cada uno fuera de lo general y común.
  + MLR
    - Modelo 1
      * Comprobación supuestos normalidad
      * Anderson test curva solo para MLR
      * Dir que es va plotjear les zones on els residuos no estabna disteibuidos normalmente.
  + PLS
  + DT
  + RF
  + XGBOOST
  + SVR

### Tipo de problema y algoritmos utilizados

En esta investigación nos enfrentamos a un problema de regresión múltiple, ya el objetivo era predecir los valores promedio de la fuerza de reacción del suelo en dirección anterior-posterior (fuerza de cizallamiento de frenado y propulsión) de la última sesión de terapia en función de las características de la primera sesión. Esto, sumado a que los conjuntos de datos estaban etiquetados, fue el motivo por el cual se utilizaron diversos algoritmos de regresión de aprendizaje automático supervisados.

Los algoritmos seleccionados para llevar a cabo las predicciones fueron la regresión linear múltiple (MLR), la regresión de mínimos cuadrados parciales o *Partial Least Squares regression* (*PLS regression*), árboles de decisión (*Decision Trees*), bosques aleatorios o *Random Forests*, eXtreme Gradient Boosting *(XGBoost)* y regresión de vectores de soporte o *Support Vector Regression (SVR)*.

### Pipeline general de entrenamiento y evaluación de los modelos

Para cada uno de los algoritmos empleados, desarrollamos funciones específicas para entrenar y evaluar el rendimiento de cada modelo en los conjuntos de entrenamiento y prueba, respectivamente. No obstante, cabe mencionar que el conjunto de datos GaitRec incluía una partición del conjunto de datos para poder crear solo los conjuntos de entrenamiento y prueba, pero no para crear el conjunto de validación. Al no disponer de un número muy elevado de observaciones en el conjunto de entrenamiento (N=180), y siguiendo las recomendaciones especificadas en la sección de “Notas de uso” del conjunto de datos GaitRec [gaitrec], realizamos la optimización de los modelos mediante distintas técnicas de validación cruzada (*cross-validation*) sólo en el conjunto de entrenamiento, manteniendo el conjunto de prueba intacto hasta la evaluación final para evitar la fuga de datos o *data leakage*.

#### Pipeline general de entrenamiento de los modelos

En la fase de entrenamiento de la mayoría de los algoritmos, usamos técnicas como la validación cruzada para realizar el ajuste de hiperparámetros (*hyperparameter* tuning) y seleccionar la mejor configuración del modelo. Una vez completado el proceso de ajuste de hiperparámetros y seleccionada la mejor configuración del modelo a partir de los resultados de la validación cruzada, entrenamos el modelo final utilizando todo el conjunto de datos de entrenamiento con los hiperparámetros seleccionados. Este proceso específico de cada algoritmo se explicará en detalle más adelante.

#### Pipeline general de evaluación de los modelos en datos no vistos

En esta subsección, presentamos el enfoque integral empleado para evaluar el rendimiento de los modelos en el conjunto de prueba, es decir, en datos no vistos.

En la fase de evaluación, evaluamos el rendimiento de los modelos en el conjunto de prueba, es decir, en los datos no vistos. Las métricas de evaluación como la raíz del error cuadrático medio (RMSE), el coeficiente de correlación de Pearson, y el coeficiente de determinación (R2) nos ofrecieron una comprensión holística de la calidad de las predicciones.

En el caso de la tercera y cuarta aproximación, al aplicar el PCA a los datos, reducimos su dimensionalidad. Sin embargo, para comprender completamente cómo se desempeñarían los modelos, era fundamental evaluar las predicciones en el espacio original de características. Es por ello por lo que evaluamos el rendimiento de los modelos después de haber realizado la transformación inversa de las predicciones al espacio de características original. Es decir, evaluamos los valores predichos dentro del espacio de características original utilizando las mismas métricas de evaluación mencionadas anteriormente. Este proceso fue fundamental para comprender el rendimiento de los modelos en las variables de salida reales, teniendo en cuenta las características del espacio original.

Además, realizamos una serie de visualizaciones que nos ayudaron a comprender e interpretar mejor los resultados obtenidos. Creamos gráficos de dispersión de los valores predichos frente a los valores reales para cada una de las 101 variable objetivo, añadiendo la línea de correlación perfecta como referencia (con una pendiente de 1 y una intersección en el origen), la línea de la regresión lineal, y la ecuación de la recta de regresión en cada gráfico. De manera similar, creamos otro gráfico de dispersión, pero en este caso con todos los puntos para comparar los valores reales y los valores predichos por los modelos de manera global. En el mismo gráfico trazamos línea de correlación perfecta que representa la relación ideal donde los valores reales y predichos son idénticos, la línea de regresión lineal que representa la relación lineal ajustada entre los valores reales y predichos, y la ecuación de la recta de regresión.

Finalmente, creamos un gráfico que resumía y presentaba de manera clara la predicción de los modelos y las métricas de evaluación empleadas. En este gráfico de líneas mostramos los valores promedio de la fuerza de reacción del suelo en dirección anterior-posterior (fuerza de cizallamiento de frenado y propulsión) en 101 puntos durante la fase de apoyo del ciclo de la marcha. Se incluyeron para su comparación los valores promedio correspondientes a la primera sesión, a la última sesión real, y a la última sesión de rehabilitación predicha por los distintos modelos. Además, se incorporó como referencia la curva que representaba los valores promedio de los controles sanos. Para tener una comprensión mayor de la calidad de las predicciones, en el mismo gráfico incluimos las métricas de evaluación como la raíz del error cuadrático medio (RMSE), el coeficiente de correlación de Pearson, y el coeficiente de determinación (R2) de los modelos.

Para complementar la evaluación de los resultados de los modelos, decidimos crear una serie de gráficos destinados a examinar y analizar la posible dependencia o relación entre el valor de las métricas de evaluación y los puntos de la curva predicha. Esto nos permitiría, por ejemplo, identificar si existían puntos de la curva más difíciles de predecir que otros, o si en los puntos críticos de la curva, como el máximo y el mínimo, podrían estar asociados con un mayor error. Para lograrlo, representamos gráficamente los valores promedio de la fuerza de reacción del suelo en dirección anterior-posterior (fuerza de cizallamiento de frenado y propulsión) en 101 puntos durante la fase de apoyo del ciclo de la marcha de última sesión de rehabilitación predicha por los distintos modelos. Entonces, asignamos colores a los puntos de la curva en base a los valores de las tres métricas de evaluación: la raíz del error cuadrático medio (RMSE), el coeficiente de correlación de Pearson, y el coeficiente de determinación (R2). Estas representaciones gráficas nos brindaron una visión más detallada y contextualizada de cómo las métricas de evaluación estaban relacionadas con los puntos individuales de la curva predicha, lo que proporcionó valiosos conocimientos sobre la calidad y las limitaciones de los modelos.

#### Validación de los modelos de regresión múltiple

Con el propósito de validar el ajuste de los modelos de regresión múltiple y saber si eran apropiados para nuestros datos, llevamos a cabo el análisis de los residuos o errores generados por los modelos.

Para realizar el análisis de los residuos, llevamos a cabo dos tipos de representaciones gráficas: el gráfico de residuos frente a los valores predichos por el modelo, y el histograma de los residuos. El gráfico de los residuos en función de los valores predichos nos permitió visualizar la dispersión de los residuos frente a los valores predichos. Es decir, identificar si los residuos se distribuían aleatoriamente alrededor del cero sin un patrón discernible, indicando un buen ajuste del modelo, o si mostraban algún tipo de patrón discernible, lo que podría indicar que el modelo no estaría capturando adecuadamente la variación en los datos. El segundo gráfico, el histograma de los residuos se utilizó para verificar si los errores seguían una distribución normal alrededor de cero, o si había alguna asimetría o sesgo en su distribución.

La validación de los modelos de regresión múltiple a través del análisis de los residuos proporcionó información sobre el ajuste de mismos a nuestros datos. Los gráficos empleados para el análisis nos permitieron evaluar tanto la dispersión como la distribución de los residuos, permitiéndonos comprender la calidad de las predicciones y la fiabilidad de los resultados obtenidos a partir de los modelos de regresión múltiple.

El proceso de evaluación descrito se encapsuló en funciones, creando un *pipeline* de evaluación apto para aplicarlo a todos los algoritmos de aprendizaje automático utilizados en el estudio y poder así, evaluar los modelos de manera automática. A continuación, se detallarán las particularidades y características propias de cada algoritmo y modelo de aprendizaje automático utilizando en este estudio.

### Regresión Linear Múltiple (MLR)

El primer algoritmo de aprendizaje automático supervisado que utilizamos fue la regresión linear múltiple (MLR). La elección de este primer algoritmo vino inspirada por el trabajo realizado por Omar A. Galarraga C. et al. descrito artículo científico titulado *Predicting Postoperative Gait in Cerebral Palsy* [Galarraga C, O. A., Vigneron, V., Dorizzi, B., Khouri, N., & Desailly, E. (2017). Predicting postoperative gait in cerebral palsy. *Gait & posture*, *52*, 45–51. <https://doi.org/10.1016/j.gaitpost.2016.11.012>] donde tratando de resolver un problema similar al nuestro, utilizaron continuación un modelo de regresión lineal múltiple para predecir simultáneamente múltiples variables cinemáticas postoperatorias.

Entrenamos y evaluamos los modelos de regresión lineal múltiple para cada una de cuatro las aproximaciones, con el propósito de explorar cómo distintos preprocesamientos podrían impactar en las predicciones de los modelos.

#### Entrenamiento y validación en el conjunto de entrenamiento

Se desarrolló una función encargada de entrenar modelos de regresión lineal múltiple utilizando la técnica de *Leave-One-Out Cross-Validation* (*LOOCV*) para evaluar el rendimiento de los modelos. Para cada una de las 101 variables objetivo, ajustamos un modelo lineal utilizando todas las observaciones excepto una, y luego evaluamos el rendimiento del modelo sobre la observación excluida. Repetimos esta evaluación para cada observación en el conjunto de entrenamiento, lo que garantizó una evaluación exhaustiva y redujo el riesgo de sobreajuste o *overfitting*. Los resultados obtenidos fueron los modelos ajustados para cada variable objetivo, las predicciones generadas durante el proceso de *LOOCV*, y los valores de la raíz del error cuadrático medio (RMSE), los valores del coeficiente de determinación (R2) y los coeficientes de correlación de Pearson, tanto para cada variable objetivo como para el modelo en global.

Entrenamos y validamos los modelos de regresión lineal múltiple para cada una de cuatro las aproximaciones. Para la primera aproximación, el modelo se entrenó con los datos de entrenamiento sin procesar. Para la segunda aproximación, aplicamos el análisis de componentes principales (PCA) solo a los datos de entrada (“X\_train”). Es decir, entrenamos el modelo con los datos de la primera sesión de terapia reducidos y los de la última sesión sin reducir (“y\_train”). Para la tercera aproximación, aplicamos el análisis de componentes principales (PCA) solo a los datos de salida (“y\_train”). Es decir, entrenamos el modelo con los datos de la primera sesión de terapia sin reducir y los de la última sesión reducidos. Para la cuarta aproximación, aplicamos el PCA tanto a los datos de entrada (“X\_train”) como a los de salida (“y\_train”). Es decir, entrenamos el modelo con los datos reducidos de la primera y última sesión de terapia. El proceso de entrenamiento para cada una de las cuatro aproximaciones se realizó de la misma forma para todos los algoritmos utilizados en el estudio, exceptuando la regresión de mínimos cuadrados parciales (PLS).

#### Evaluación en el conjunto de prueba

En esta etapa, evaluamos la calidad y el rendimiento de los modelos de regresión lineal múltiple entrenados en el conjunto de entrenamiento sobre datos no vistos, es decir, el conjunto de prueba.

Utilizando el *pipeline* de evaluación anteriormente descrito, obtuvimos de manera automática las métricas de evaluación como la raíz del error cuadrático medio (RMSE), el coeficiente de correlación de Pearson, y el coeficiente de determinación (R2) globales del modelo. Además, en el caso de la tercera y cuarta aproximación, utilizando las mismas métricas evaluamos las predicciones en el espacio original después de la transformación inversa para comprender el rendimiento de los modelos en las variables reales.

Visualizamos los gráficos de dispersión de los valores predichos frente a los valores reales para cada una de las 101 variable objetivo, así como el gráfico de dispersión con todos los puntos para comparar los valores reales y los valores predichos por los modelos de manera global. Mostramos también el gráfico que resumía y presentaba de manera clara la predicción de los modelos y las métricas de evaluación empleadas. Para complementar la evaluación de los resultados de los modelos, representamos los distintos gráficos destinados a examinar y analizar la posible dependencia o relación entre el valor de las métricas de evaluación y los puntos de la curva predicha.

Finalmente, validamos el ajuste de los modelos de regresión múltiple mediante el análisis de los residuos o errores generados por los modelos. Para ello representamos los residuos frente a los valores predichos por el modelo, y el histograma de los residuos.

Evaluamos los modelos de regresión lineal múltiple en el conjunto de prueba para cada una de cuatro las aproximaciones. Para la primera aproximación, evaluamos el rendimiento del modelo en los datos de prueba sin procesar. Para la segunda aproximación, aplicamos el análisis de componentes principales (PCA) solo a los datos de entrada (“X\_test”). Es decir, evaluamos el rendimiento del modelo en los datos de prueba de la primera sesión de terapia reducidos y en los de la última sesión sin reducir (“y\_test”). Para la tercera aproximación, aplicamos el análisis de componentes principales (PCA) solo a los datos de salida (“y\_test”). Es decir, evaluamos el rendimiento del modelo en los datos de prueba de la primera sesión de terapia sin reducir y en los de la última sesión reducidos. Para la cuarta aproximación, aplicamos el PCA tanto a los datos de entrada (“X\_test”) como a los de salida (“y\_test”). Es decir, evaluamos el rendimiento del modelo en los datos de prueba reducidos de la primera y última sesión de terapia. En el caso de la tercera y cuarta aproximación, después de obtener las predicciones de los modelos de aprendizaje automático, llevamos a cabo una transformación inversa de esas predicciones. Esta transformación nos permitió regresar al espacio original de las variables objetivo. En estas dos aproximaciones específicas, el análisis de residuos se realizó utilizando las predicciones que habían sido reconstruidas al espacio original de características.

El proceso de evaluación en cada una de las cuatro aproximaciones se realizó de la misma forma para todos los algoritmos utilizados en el estudio, exceptuando la regresión de mínimos cuadrados parciales (PLS).

#### Comprobación de los supuestos

Para verificar si podíamos usar los modelos de regresión lineal múltiple y garantizar que fuesen válidos y fiables, comprobamos los supuestos más importantes: el supuesto de linealidad, de independencia de errores o residuos, de normalidad de los residuos, y de homocedasticidad.

Para comprobar el supuesto de linealidad, representamos gráficamente los residuos frente a los valores predichos para cada variable objetivo. Esto nos permitió visualizar la relación entre los residuos y las predicciones y evaluar si el supuesto de linealidad se cumplía para cada variable objetivo. Adicionalmente, creamos gráficos conjuntos (*joint plots)* entre cada predictor y variable objetivo, mostrando la relación existente entre ellos.

Para comprobar el supuesto de independencia de los residuos, examinamos el gráfico de residuos frente a los valores predichos para cada variable objetivo añadiendo el resultado la prueba estadística de Durbin-Watson en cada uno. Esto nos permitió detectar la posible presencia de autocorrelación (dependencia de los errores) en los residuos.

Para comprobar el supuesto de normalidad de los residuos utilizamos la prueba estadística de Anderson-Darling. Con esta prueba determinamos si los residuos seguían una distribución normal al nivel de significación del 5% para cada variable objetivo.

Finalmente, para comprobar el supuesto de homocedasticidad en los residuos, utilizamos la prueba estadística de Breusch-Pagan. Mediante el *p-value*, esta prueba nos permitió comprobar si varianza de los residuos era constante a lo largo de los valores predichos.

Después de comprobar los supuestos, en la fase de evaluación de la regresión lineal múltiple se crearon dos gráficos adicionales que representaban los valores promedio de la fuerza de reacción del suelo en dirección anterior-posterior (fuerza de cizallamiento de frenado y propulsión) en 101 puntos durante la fase de apoyo del ciclo de la marcha de última sesión de rehabilitación predicha por los distintos modelos. En el primer gráfico, asignamos colores a los puntos de la curva en base a los valores de la prueba estadística de Anderson-Darling. En el segundo gráfico coloreamos de rojo los puntos de la curva predicha dónde los residuos no seguían una distribución normal.

### Regresión de mínimos cuadrados parciales (*PLS regression*)

El segundo algoritmo de aprendizaje automático supervisado que utilizamos fue la regresión de mínimos cuadrados parciales (PLS regression). Elegimos el PLS debido a que crea componentes latentes que reduzcan la dimensionalidad manteniendo información relevante. Por tanto, es útil cuando se trabaja con conjuntos de datos donde hay muchas variables predictoras y existe una posible multicolinealidad. Además, puede manejar tanto relaciones lineales como no lineales en los datos.

Entrenamos y evaluamos los modelos de regresión de mínimos cuadrados parciales para cada una de cuatro las aproximaciones, con el propósito de explorar cómo distintos preprocesamientos podrían impactar en las predicciones de los modelos.

#### Entrenamiento y validación en el conjunto de entrenamiento

Desarrollamos una función encargada de entrenar modelos de regresión de mínimos cuadrados parciales y evaluar su rendimiento en el conjunto de entrenamiento. Aunque generalmente se recomienda escalar las variables antes de llevar a cabo un PLS, en nuestro caso decidimos no hacerlo debido a que los datos postprocesados proporcionados por GaitRec habían sido normalizados previamente. Para encontrar el número óptimo de componentes o variables latentes para el modelo PLS, realizamos el ajuste de hiperparámetros (*hyperparameter tuning*), que se llevó a cabo mediante *GridSearchCV. Se* utilizóuna validación cruzada de 10 iteraciones (*10-fold cross-validation*) y la métrica de puntuación (*scoring)* se estableció en la raíz del error cuadrático medio negativo (RMSE) para facilitar la minimización del error. Representamos gráficamente los valores del RMSE correspondientes a diferentes números de componentes, con el mejor número de componentes señalado, lo que ayudó a visualizar y tener una referencia de la elección óptima. Sin embargo, con el objetivo de evitar aumentar la complejidad del modelo, elegimos el número de componentes en el que la diferencia entre un componente y el siguiente no superara el 2%. Una vez seleccionado el número óptimo de componentes, volvimos a entrenar el modelo de PLS utilizando la configuración optima de 8 componentes. Esto fue un paso esencial para obtener el modelo final con la configuración óptima. A continuación, realizamos la validación cruzada con 10 iteraciones (*10-fold cross-validation*) y obtuvimos las predicciones utilizando el modelo entrenado. Finalmente, los resultados obtenidos para el conjunto de entrenamiento fueron el modelo PLS, los valores de la raíz del error cuadrático medio (RMSE) y del coeficiente de determinación (R2), y los coeficientes de correlación de Pearson para el modelo en global.

En el caso del PLS, seguimos un enfoque diferente en comparación con las aproximaciones previamente explicadas. Esto se debe a que el PLS es una técnica de reducción de la dimensionalidad supervisada más avanzada que el PCA, ya que considera de manera integrada la relación entre las variables predictoras (X) y las variables objetivo (y). El PLS crea variables latentes que maximizan la covarianza entre las matrices X e y. A diferencia del PCA, que se enfoca únicamente en reducir la dimensionalidad de las variables predictoras, el PLS toma en cuenta tanto las características de la entrada como la relación con la salida. Esta característica lo convierte en una técnica más completa y con un mayor potencial en términos de rendimiento predictivo. En este sentido, el PLS podría considerarse una evolución más avanzada de la segunda aproximación.

#### Evaluación en el conjunto de prueba

En esta etapa, evaluamos la calidad y el rendimiento del modelo de regresión de mínimos cuadrados parciales entrenado en el conjunto de entrenamiento sobre datos no vistos, es decir, sobre el conjunto de prueba.

Utilizando el *pipeline* de evaluación anteriormente descrito, obtuvimos de manera automática las métricas de evaluación como la raíz del error cuadrático medio (RMSE), el coeficiente de correlación de Pearson, y el coeficiente de determinación (R2) globales del modelo. Visualizamos los gráficos de dispersión de los valores predichos frente a los valores reales para cada una de las 101 variable objetivo, así como el gráfico de dispersión con todos los puntos para comparar los valores reales y los valores predichos por los modelos de manera global. Mostramos también el gráfico que resumía y presentaba de manera clara la predicción de los modelos y las métricas de evaluación empleadas. Para complementar la evaluación de los resultados de los modelos, representamos los distintos gráficos destinados a examinar y analizar la posible dependencia o relación entre el valor de las métricas de evaluación y los puntos de la curva predicha. Finalmente, validamos el ajuste de los modelos de regresión múltiple mediante el análisis de los residuos o errores generados por los modelos. Para ello representamos los residuos frente a los valores predichos por el modelo, y el histograma de los residuos.

### Árboles de decisión

El tercer algoritmo de aprendizaje automático supervisado que utilizamos fueron los árboles de decisión o *Decision Trees*. Este algoritmo se utilizó como un punto de partida o una primera aproximación a las técnicas de *ensamble learning* más avanzadasutilizadas posteriormente, como los bosques aleatorios (*Random Forest)* y *Gradient Boosting*.

Los árboles de decisión, que son algoritmos no lineales de aprendizaje automático, podrían ser apropiados para nuestra investigación ya que pueden resolver problemas de regresión, las predicciones generadas son fáciles de interpretar y explicar, y son capaces de modelar relaciones complejas entre las variables de entrada y predecir múltiples variables de salida simultáneamente. No obstante, a diferencia de algoritmos lineales como la regresión lineal múltiple o el PLS, los árboles de decisión son propensos a la alta varianza.

Entrenamos y evaluamos los modelos de regresión lineal múltiple para cada una de cuatro las aproximaciones, con el propósito de explorar cómo distintos preprocesamientos podrían impactar en las predicciones de los modelos.

#### Entrenamiento y validación en el conjunto de entrenamiento

Desarrollamos una función encargada de entrenar los árboles de decisión y evaluar su rendimiento en el conjunto de entrenamiento.

Inicializamos el árbol de regresión y definimos una serie de hiperparámetros para ajustar el modelo.

Los hiperparámetros que incluimos en el proceso de *hyperparameter tuning* fueron el máximo de profundidad del árbol (*max\_depth*), el número mínimo de muestras requeridas para dividir un nodo interno (*min\_samples\_split*) y el número mínimo de muestras requeridas en un nodo hoja (*min\_samples\_leaf*).

Los rangos de valores para los hiperparámetros se muestran en la Figura X. Utilizamos la técnica de *GridSearchCV* para explorar diferentes las combinaciones de hiperparámetros. Esta búsqueda se realizó utilizando una validación cruzada de 10 iteraciones para evitar el sobreajuste, y la métrica de puntuación (*scoring)* se estableció en la raíz del error cuadrático medio negativo (RMSE) para facilitar la minimización del error.

Identificamos los mejores hiperparámetros que generaron el menor RMSE en la validación cruzada. Estos parámetros optimizados se utilizaron para entrenar el modelo final en el conjunto de entrenamiento.

A continuación, realizamos la validación cruzada con 10 iteraciones (*10-fold cross-validation*) utilizando el modelo optimizado y obtuvimos las predicciones de las variables de salida.

Finalmente, los resultados obtenidos para el conjunto de entrenamiento fueron el modelo de árbol de decisión, los valores de la raíz del error cuadrático medio (RMSE) y del coeficiente de determinación (R2), y los coeficientes de correlación de Pearson para el modelo en global.

A screenshot of a computer

Description automatically generated

Figura X. Rangos de valores para los hiperparámetros de los árboles de decisión.

Entrenamos y validamos el modelo de árbol de decisión para cada una de cuatro las aproximaciones, de la misma forma que con el algoritmo de regresión lineal múltiple. Como se mencionó en apartados anteriores, el proceso de entrenamiento para cada una de las cuatro aproximaciones se realizó de la misma forma para todos los algoritmos utilizados en el estudio, exceptuando la regresión de mínimos cuadrados parciales (PLS).

#### Evaluación en el conjunto de prueba

En esta etapa, evaluamos la calidad y el rendimiento del modelo de árbol de decisión entrenado en el conjunto de entrenamiento sobre datos no vistos, es decir, sobre el conjunto de prueba. Como se comentó en apartados anteriores, el proceso de evaluación se realizó de la misma forma para todos los algoritmos, exceptuando la regresión de mínimos cuadrados parciales (PLS).

### Random Forest

El cuarto algoritmo de aprendizaje automático supervisado que utilizamos fue el bosque aleatorio o *random forest*. Dado que ya habíamos explorado los árboles de decisión, optamos llevar nuestra investigación a un nivel superior utilizando métodos de *ensamble learning* como el *random forest*. Los bosques aleatorios están formados por muchos árboles de decisión individuales, y utilizan la agregación *bootstrap* para combinar estos árboles y alcanzar una solución mucho mejor que la proporcionada por cualquiera de los árboles individuales.

#### Entrenamiento y validación en el conjunto de entrenamiento

Desarrollamos una función encargada de entrenar el modelo de *random forest* y evaluar su rendimiento en el conjunto de entrenamiento.

Inicializamos el bosque aleatorio y definimos una serie de hiperparámetros para ajustar el modelo.

Los hiperparámetros que utilizamos en el proceso de *hyperparameter tuning* incluyeron el número de árboles en el bosque aleatorio (*n\_estimators*), el número de características a considerar en cada división (*max\_features*), la profundidad máxima de cada árbol (*max\_depth*), el número mínimo de muestras necesarias para dividir un nodo (*min\_samples\_split*), y el número mínimo de muestras necesarias en cada nodo hoja (*min\_samples\_leaf*). Los rangos de valores para los hiperparámetros se muestran en la Figura X.

Para explorar la amplia gama de combinaciones de hiperparámetros de manera más eficiente, optamos por utilizar la técnica de *RandomizedSearchCV* en lugar de *GridSearchCV* para el ajuste de hiperparámetros en el modelo de *random forest*. Dado que GridSearchCV evalúa todas las combinaciones posibles, se volvió computacionalmente costoso en nuestro conjunto de datos. *RandomizedSearchCV*, por otro lado, realiza una búsqueda aleatoria de hiperparámetros, lo que redujo significativamente el tiempo de ejecución y permitió la exploración de diversas configuraciones, lo cual fue esencial para lograr un equilibrio entre rendimiento y eficiencia.

La búsqueda se realizó utilizando una validación cruzada de 10 iteraciones para evitar el sobreajuste, y la métrica de puntuación (*scoring)* se estableció en la raíz del error cuadrático medio negativo (RMSE) para facilitar la minimización del error.

Identificamos los mejores hiperparámetros que generaron el menor RMSE en la validación cruzada. Estos parámetros optimizados se utilizaron para entrenar el modelo final en el conjunto de entrenamiento. A continuación, realizamos la validación cruzada con 10 iteraciones (*10-fold cross-validation*) utilizando el modelo optimizado y obtuvimos las predicciones de las variables de salida.

Finalmente, los resultados obtenidos para el conjunto de entrenamiento fueron el modelo de *random forest*, los valores de la raíz del error cuadrático medio (RMSE) y del coeficiente de determinación (R2), y los coeficientes de correlación de Pearson para el modelo en global.

Entrenamos y validamos el modelo de bosque aleatorio para cada una de cuatro las aproximaciones, de la misma forma que el resto de los algoritmos utilizados en el estudio, exceptuando la regresión de mínimos cuadrados parciales (PLS).

A computer screen shot of a program code

Description automatically generated

Figura X. Rangos de valores para los hiperparámetros de los árboles aleatorios.

#### Evaluación en el conjunto de prueba

Finalmente, evaluamos la calidad y el rendimiento del modelo de *random forest* entrenado en el conjunto de entrenamiento sobre datos no vistos, es decir, sobre el conjunto de prueba. Como se comentó en apartados anteriores, el proceso de evaluación se realizó de la misma forma para todos los algoritmos, exceptuando la regresión de mínimos cuadrados parciales (PLS).

### Extreme Gradient Boosting (XGBoost)

El quinto algoritmo de aprendizaje automático supervisado que utilizamos fue el *eXtreme Gradient Boosting (XGBoost)*.

La elección de implementar el *XGBoost* se fundamenta en una estrategia progresiva y evolutiva en relación con los algoritmos basados en árboles que utilizamos previamente. Empezamos con los árboles de decisión, y luego seguimos con los bosques aleatorios. Por tanto, en este contexto *XGBoost* es el próximo paso. Esta técnica de *ensamble learning* también se basa en árboles, pero introduce mejoras significativas en términos de eficiencia y rendimiento. *XGBoost* optimiza la creación de árboles individuales mediante el uso de *gradient boosting* y técnicas de regularización. Esto ayuda a mitigar el sobreajuste o *overfitting* y a obtener predicciones más precisas en comparación con los algoritmos anteriores.

#### Entrenamiento y validación en el conjunto de entrenamiento

Para la implementación de *eXtreme Gradient Boosting* en nuestra investigación, desarrollamos una función encargada de entrenar el modelo de *XGBoost* y evaluar su rendimiento en el conjunto de entrenamiento. Entrenamos un único modelo con *hyperparameter tuning* y lo utilizamos para predecir todas las 101 variables objetivo simultáneamente. Este enfoque simplificó el proceso de modelado, pero cabía la posibilidad de sacrificar algo de interpretabilidad.

Inicializamos el *XGBoostRegressor* y definimos una serie de hiperparámetros para ajustar el modelo. Los hiperparámetros que utilizamos en el proceso de *hyperparameter tuning* incluyeron el número de estimadores o árboles en el ensamble *(n\_estimators)*, la tasa de aprendizaje *(learning\_rate)*, el peso mínimo requerido en un nodo hijo para continuar dividiendo el nodo *(min\_child\_weight)*, la profundidad máxima de cada árbol *(max\_depth)*, la proporción de muestras utilizadas en cada iteración *(subsample)*, la proporción de características utilizadas en cada árbol *(colsample\_bytree)*, la proporción de características utilizadas en cada nivel del árbol *(colsample\_bylevel)*, el parámetro de regularización L2 o Ridge *(reg\_lambda)*, el parámetro de regularización L1 o Lasso *(reg\_alpha)*, y el parámetro que controla cuánta reducción de pérdida es necesaria para realizar una partición *(gamma)*. Los rangos de valores para los hiperparámetros se muestran en la Figura X.

Para explorar la amplia gama de combinaciones de hiperparámetros de manera más eficiente, optamos por utilizar la técnica de *RandomizedSearchCV*. Este enfoque nos permitió obtener combinaciones óptimas en menos tiempo.

La búsqueda se realizó utilizando una validación cruzada de 5 iteraciones para evitar el sobreajuste, y la métrica de puntuación (*scoring)* se estableció en la raíz del error cuadrático medio negativo (RMSE) para facilitar la minimización del error.

Identificamos los mejores hiperparámetros que generaron el menor RMSE en la validación cruzada. Estos parámetros optimizados se utilizaron para entrenar el modelo final en el conjunto de entrenamiento. A continuación, realizamos la validación cruzada con 10 iteraciones (*10-fold cross-validation*) utilizando el modelo optimizado y obtuvimos las predicciones de las variables de salida.

Finalmente, al igual que con los algoritmos previos, los resultados obtenidos para el conjunto de entrenamiento fueron el modelo de *XGBoost*, los valores de la raíz del error cuadrático medio (RMSE) y del coeficiente de determinación (R2), y los coeficientes de correlación de Pearson para el modelo en global.

Entrenamos y validamos el modelo de *XGBoost* para cada una de cuatro las aproximaciones, de la misma forma que el resto de los algoritmos utilizados en el estudio, exceptuando la regresión de mínimos cuadrados parciales (PLS).

A screenshot of a computer program

Description automatically generated

Figura X. Rangos de valores para los hiperparámetros del eXtreme Gradient Boosting (XGBoost).

#### Evaluación en el conjunto de prueba

Finalmente, evaluamos la calidad y el rendimiento del modelo de *XGBoost* entrenado en el conjunto de entrenamiento sobre datos no vistos, es decir, sobre el conjunto de prueba. Como se comentó en apartados anteriores, el proceso de evaluación se realizó de la misma forma para todos los algoritmos, exceptuando la regresión de mínimos cuadrados parciales (PLS).

### Support Vector Regression (SVR)

El sexto (y último) algoritmo de aprendizaje automático supervisado que utilizamos fue la regresión de vectores de soporte o *Support Vector Regression* *(SVR).*

Teniendo en cuenta los algoritmos utilizados hasta ahora, la regresión de vectores de soporte proporcionó una perspectiva adicional a nuestro estudio.

#### Entrenamiento y validación en el conjunto de entrenamiento

Para la implementación de regresión de vectores de soporte en nuestra investigación, desarrollamos una función encargada de entrenar los modelos de *SVR* y evaluar su rendimiento en el conjunto de entrenamiento.

Sin embargo, cabe destacar que no todos los algoritmos de regresión admiten la regresión múltiple. La regresión de vectores de soporte es un ejemplo de ello. Por tanto, tuvimos que desarrollar un modelo independiente para cada valor numérico a predecir, lo que se conoce como enfoque o aproximación directa (*direct approach*). Aunque esto supone que las salidas son independientes entre sí, lo que podría no ser una suposición correcta.

Implementamos un proceso de entrenamiento y evaluación utilizando el concepto de *MultiOutputRegressor*. Este enfoque nos permitió manejar simultáneamente múltiples variables objetivo.

Realizamos *hyperparameter tuning* con el objetivo de encontrar la configuración óptima para el modelo *SVR*. Utilizamos la técnica de *RandomizedSearchCV* para explorar diversas combinaciones de hiperparámetros. Los hiperparámetros que utilizamos fueron el parámetro de regularización que controla la penalización por errores *(C)*, la tolerancia del error *(epsilon)*, la función de *kernel* utilizada en el modelo SVR *(kernel)*, el coeficiente del kernel, que afecta la forma de la función de decisión *(gamma)*, y el grado del *kernel* polinómico *(degree)*. Los rangos de valores para los hiperparámetros se muestran en la Figura X.

Para explorar la amplia gama de combinaciones de hiperparámetros de manera eficiente, optamos por utilizar la técnica de *RandomizedSearchCV*.

La búsqueda se realizó utilizando una validación cruzada de 5 iteraciones para evitar el sobreajuste, y la métrica de puntuación (*scoring)* se estableció en la raíz del error cuadrático medio negativo (RMSE) para facilitar la minimización del error.

Identificamos los mejores hiperparámetros que generaron el menor RMSE en la validación cruzada. Estos parámetros optimizados se utilizaron para entrenar el modelo final en el conjunto de entrenamiento. A continuación, realizamos la validación cruzada con 10 iteraciones (*10-fold cross-validation*) utilizando el modelo optimizado y obtuvimos las predicciones de las variables de salida.

Finalmente, al igual que con los algoritmos previos, los resultados obtenidos para el conjunto de entrenamiento fueron el modelo de *SVR*, los valores de la raíz del error cuadrático medio (RMSE) y del coeficiente de determinación (R2), y los coeficientes de correlación de Pearson para el modelo en global.

Entrenamos y validamos el modelo de regresión de vectores de soporte para cada una de cuatro las aproximaciones, de la misma forma que el resto de los algoritmos utilizados en el estudio, exceptuando la regresión de mínimos cuadrados parciales (PLS).

A computer code with colorful text

Description automatically generated

Figura X. Rangos de valores para los hiperparámetros de la regresión de vectores de soporte (SVR).

#### Evaluación en el conjunto de prueba

Finalmente, evaluamos la calidad y el rendimiento del modelo de regresión de vectores de soporte entrenado en el conjunto de entrenamiento sobre datos no vistos, es decir, sobre el conjunto de prueba. Como se comentó en apartados anteriores, el proceso de evaluación se realizó de la misma forma para todos los algoritmos, exceptuando la regresión de mínimos cuadrados parciales (PLS).

# Resultados

Resultados en pasado

* Número de pacientes por grupo o clase

A graph of different colored bars

Description automatically generated

* Número de pacientes únicos para cada subclase

A graph of different colored bars

Description automatically generated

* distribución de pacientes por número de sesiones

A graph of a number of sessions

Description automatically generated

* + observamos que la mayoría de pacientes han realizado 2, 3 o 4 sesiones.
* 50 sujetos con más *trials*

A graph of a number of objects

Description automatically generated with medium confidence

* 50 sujetos con menos *trials:*

A graph with numbers and a bar

Description automatically generated with medium confidence

* Distribución de los sujetos por número de *trials.*

A graph of a number of trials

Description automatically generated

**Agrupar por grupo y número de sesiones:**

* Debido a que la mayoría de sujetos tienen de 2 a 4 sesiones, hemos elegido este rango: nos quedamos con los pacientes que tengan 2, 3 o 4 sesiones.

**Plot group 1st session mean + group last session mean + HC mean**

A graph showing the results of a performance

Description automatically generated with medium confidence**A graph showing the difference between the first and last session mean and the first session mean

Description automatically generated**

**A graph showing different colored lines

Description automatically generated**A graph of a graph showing different colored lines

Description automatically generated with medium confidence**A graph with red blue and green lines

Description automatically generatedA graph of a graph showing different colored lines

Description automatically generated with medium confidenceA graph showing the difference between the same line

Description automatically generated with medium confidenceA graph showing the difference between a number of objects

Description automatically generated with medium confidenceA graph showing the difference between the same number of classes

Description automatically generated with medium confidenceA graph with red and blue lines

Description automatically generated**

## EDA

A chart of blue and white shapes

Description automatically generated with medium confidence

A red and blue triangle

Description automatically generated

## Regresión Linear Múltiple

### Primera aproximación

# Discusiones

Discusiones en presente

## Limitaciones del estudio

# Consideraciones éticas

# Coste económico

# Conclusiones

# Vías abiertas

# Referencias

1. Example 1
2. Example 2

[*Eric J Topol. 2019. High-performance medicine: The convergence of human and artificial intelligence. Nature Medicine 25, 1 (2019), 44–56.* [*https://doi.org/10.1038/s41591-018-0300-7*](https://doi.org/10.1038/s41591-018-0300-7)]

[*Joana Figueiredo, Cristina P. Santos, and Juan C. Moreno. 2018. Automatic recognition of gait patterns in human motor disorders using machine learning: A review. Medical Engineering and Physics 53 (2018), 1–12.* [*https://doi.org/10.1016/j.medengphy.2017.12.006*](https://doi.org/10.1016/j.medengphy.2017.12.006)]

[*Wolfgang I Schöllhorn. 2004. Applications of artificial neural nets in clinical biomechanics. Clinical Biomechanics 19, 9 (2004), 876–898.* [*https://doi.org/10.1016/j.clinbiomech.2004.04.005*](https://doi.org/10.1016/j.clinbiomech.2004.04.005)]

[gaitrec] Horsak, B., Slijepcevic, D., Raberger, AM. *et al.* GaitRec, a large-scale ground reaction force dataset of healthy and impaired gait. *Sci Data* **7**, 143 (2020). <https://doi.org/10.1038/s41597-020-0481-z>

[*GaitRec-Net: A Deep Neural Network for Gait Disorder Detection Using Ground Reaction Force*] Pandey, C., Roy, D., Poonia, R., Altameem, A., Nayak, S., Verma, A., & Saudagar, A. (2022). GaitRec-Net: A Deep Neural Network for Gait Disorder Detection Using Ground Reaction Force. PPAR Research, 2022. <https://doi.org/10.1155/2022/9355015>.

[*Hong-yin Lau, Kai-yu Tong, and Hailong Zhu. 2009. Support vector machine for classification of walking conditions of persons after stroke with dropped foot. Human Movement Science 28, 4 (Aug. 2009), 504–514.* [*https://doi.org/10.1016/j.humov.2008.12.003*](https://doi.org/10.1016/j.humov.2008.12.003)]

[*Leen Van Gestel, Tinne De Laet, Enrico Di Lello, Herman Bruyninckx, Guy Molenaers, Anja Van Campenhout, Erwin Aertbeliën, Mike Schwartz, Hans Wambacq, Paul De Cock, and Kaat Desloovere. 2011. Probabilistic gait classification in children with cerebral palsy: A Bayesian approach. Research in Developmental Disabilities 32, 6 (Nov. 2011), 2542–2552. https://doi.org/10.1016/j.ridd.2011.07.004*]

[*Corina Nüesch, Victor Valderrabano, Cora Huber, Vinzenz von Tscharner, and Geert Pagenstert. 2012. Gait patterns of asymmetric ankle osteoarthritis patients. Clinical Biomechanics 27, 6 (July 2012), 613–618.* [*https://doi.org/10.1016/j.clinbiomech.2011.12.016*](https://doi.org/10.1016/j.clinbiomech.2011.12.016)]

[*Murad Alaqtash, Thompson Sarkodie-Gyan, Huiying Yu, Olac Fuentes, Richard Brower, and Amr Abdelgawad. 2011. Automatic classification of pathological gait patterns using ground reaction forces and machine learning algorithms. In 2011 Annual International*

*Conference of the IEEE Engineering in Medicine and Biology Society (EMBS). IEEE, 453–457.* [*https://doi.org/10.1109/IEMBS.2011*](https://doi.org/10.1109/IEMBS.2011)*. 6090063*]

[*Ferdous Wahid, Rezaul K Begg, Chris J Hass, Saman Halgamuge, and David C Ackland. 2015. Classification of Parkinson’s disease gait using spatial-temporal gait features. IEEE Journal of Biomedical and Health Informatics 19, 6 (2015), 1794–1802.*]

[*Djordje Slijepcevic, Matthias Zeppelzauer, Anna-Maria Gorgas, Caterine Schwab, Michael Schüller, Arnold Baca, Christian Breiteneder, and Brian Horsak. 2017. Automatic classification of functional gait disorders. IEEE Journal of Biomedical and Health Informatics 22, 5 (2017), 1653–1661. https://doi.org/10.1109/JBHI.2017.2785682*]

[O. A. Galarraga C., V. Vigneron, B. Dorizzi, N. Khouri & E. Desailly (2015) Predicting postoperative knee flexion during gait of cerebral palsy children, Computer Methods in Biomechanics and Biomedical Engineering, 18:sup1, 1940-1941, DOI: 10.1080/10255842.2015.1070583]

[Galarraga, Omar & Vigneron, Vincent & Dorizzi, Bernadette & Khouri, Nejib & Desailly, Eric. (2015). Estimation of Postoperative Knee Flexion at Initial Contact of Cerebral Palsy Children using Neural Networks. ICPRAM 2015 - 4th International Conference on Pattern Recognition Applications and Methods, Proceedings. 2. 10.5220/0005286503380342.]

[Galarraga C, O. A., Vigneron, V., Dorizzi, B., Khouri, N., & Desailly, E. (2017). Predicting postoperative gait in cerebral palsy. *Gait & posture*, *52*, 45–51. <https://doi.org/10.1016/j.gaitpost.2016.11.012>]

[pandas] McKinney, W., & others. (2010). Data structures for statistical computing in python. In *Proceedings of the 9th Python in Science Conference* (Vol. 445, pp. 51–56).

[numpy] Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., … Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, *585*, 357–362. <https://doi.org/10.1038/s41586-020-2649-2>

[matplotlib] Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science &amp; Engineering*, *9*(3), 90–95.

[seaborn] Waskom, M., Botvinnik,                 Olga, O&#x27;Kane,                 Drew, Hobson,                 Paul, Lukauskas,                 Saulius, Gemperline,                 David C, … Qalieh,                 Adel. (2017). *mwaskom/seaborn: v0.8.1 (September 2017)*. Zenodo. <https://doi.org/10.5281/zenodo.883859>

[scipy] Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., … SciPy 1.0 Contributors. (2020). SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, *17*, 261–272. <https://doi.org/10.1038/s41592-019-0686-2>

[statsmodels] Seabold, S., & Perktold, J. (2010). statsmodels: Econometric and statistical modeling with python. In *9th Python in Science Conference*.

[scikit-learn] Pedregosa, F., Varoquaux, Ga"el, Gramfort, A., Michel, V., Thirion, B., Grisel, O., … others. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*(Oct), 2825–2830.

[xgboost] Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). New York, NY, USA: ACM. <https://doi.org/10.1145/2939672.2939785>

[git] Chacon, S., & Straub, B. (2014). *Pro git*. Apress.

[github] github. (2020). *GitHub*. Retrieved from <https://github.com/>

[*Mullineaux, D. R., Milner, C. E., Davis, I. S. & Hamill, J. Normalization of ground reaction forces. J. Appl. Biomech. 22, 230–233 (2006).*]

[*Mullineaux, D. R., Milner, C. E., Davis, I. S. & Hamill, J. Normalization of ground reaction forces. J. Appl. Biomech. 22, 230–233 (2006).*]*,* [*Helwig, N. E., Hong, S., Hsiao-Wecksler, E. T. & Polk, J. D. Methods to temporally align gait cycle data. J. Biomech. 44, 561–566 (2011).*].

[*Sangeux, M. & Polak, J. A simple method to choose the most representative stride and detect outliers. Gait Posture 41, 726–730 (2015).*]

# Apéndice

## Apéndice 1. Repositorio del proyecto

El código fuente desarrollado para este proyecto está disponible en el repositorio GitHub: https://github.com/bernatsort/Therapy-Progress-Gait-Prediction.git