Los medicamentos del futuro: Análisis y predicción del tiempo de desarrollo de nuevos fármacos mediante técnicas de aprendizaje automático

Miquel Ribas Portella

Machine learning applied to Clinical Trials

Máster Universitario en Ciencia de Datos

Nombre del director/a de TF:
Susana Pérez Álvarez

Nombre del/de la PRA:
Laia Subirats Maté

12 de octubre de 2025

Universitat Oberta
de Catalunya

Ficha Del Trabajo Final

Título del trabajo:	Los medicamentos del futuro: Análisis y predicción del tiempo de desarrollo de nuevos fármacos mediante técnicas de aprendizaje automático
Nombre del autor/a:	Miquel Ribas Portella
Nombre del director/a de TF:	Susana Pérez Álvarez
Nombre del/de la PRA:	Laia Subirats Maté
Fecha de entrega:	12 de octubre de 2025
Titulación o programa:	Máster Universitario en Ciencia de Datos
Área del trabajo final:	Machine learning applied to Clinical Trials
Idioma del trabajo:	Castellano
Palabras clave:	Aprendizaje automático; Machine learning; Ensayos clínicos; Clinical trials; Modelos predictivos; Predictive modeling; Datos clínicos; Clinical data; Desarrollo de fármacos; Drug development

Resumen del trabajo

Los ensayos clínicos constituyen una etapa esencial en el desarrollo de nuevos medicamentos, tanto por su complejidad como por los elevados costes y tiempos asociados. Este trabajo tiene como objetivo aplicar técnicas de aprendizaje automático para analizar y predecir la duración de las fases de los ensayos clínicos, con el fin de estimar el tiempo total de desarrollo de nuevos fármacos. Se emplearán modelos predictivos basados en datos clínicos públicos identificando patrones relacionados con la duración, tipo de estudio, área terapéutica, y resultados previos.

Se espera demostrar el potencial del aprendizaje supervisado como herramienta de apoyo a la planificación estratégica del desarrollo farmacéutico, contribuyendo a optimizar los recursos (tiempo y costes) y a anticipar qué medicamentos podrían llegar al mercado en los próximos años.

Abstract

Clinical trials represent an essential stage in the development of new drugs, both because of their complexity and the high costs and time involved.
The aim of this work is to apply machine learning techniques to analyze and predict the duration of the phases of the clinical trials, in order to estimate the overall drug development time. Predictive models based on publicly available clinical data will be employed, identifying patterns related to duration, type of study, therapeutic area and previous outcomes.
The study aims to demonstrate the potential of supervised learning as a strategic planning support tool during drug development, helping to optimize resources (time, costs) and anticipate which drugs could reach the market in the following years.

Índice general

Índice de figuras

Índice de cuadros

Capítulo 1

Introducción

1.1. Contexto y justificación del trabajo

El desarrollo de fármacos depende del éxito de los ensayos clínicos, un proceso caracterizado por su complejidad, duración y alto coste económico. Muchos ensayos fracasan en fases avanzadas (III, IV) debido a una selección inadecuada de pacientes o estimaciones imprecisas de la eficacia terapéutica de los fármacos.

El uso de técnicas de aprendizaje automático en el ámbito biomédico permite nuevas oportunidades de optimizar estos procesos mediante la integración de grandes volúmenes de datos clínicos. Los modelos predictivos buscan anticipar resultados, identificar patrones en la duración de las fases clínicas y ajustar parámetros del diseño experimental, con el fin de reducir la tasa de fracaso de los ensayos. El diseño de un ensayo clínico es una fase crítica, ya que de el dependen la validez de los resultados y la eficiencia del proceso.

El proyecto resulta relevante porque contribuye a la transformación digital del sector farmacéutico, mediante el análisis temporal de los ensayos clínicos disponibles en ClinicalTrials.gov[1], con métodos avanzados de aprendizaje automático, y contribuye a mejorar la planificación estratégica y optimizar recursos (tiempo y costes).

1.2. Explicación de la motivación personal

Mi motivación se basa en el interés profesional de la combinación entre la ciencia de datos y el sector farmacéutico, donde trabajo actualmente como consultor senior.

He podido observar de cerca los retos que enfrenta la industria en el diseño y gestión de ensayos clínicos y personalmente veo potencial en los modelos de aprendizaje automático para optimizar estos procesos.

Este TFM me permitirá aplicar y consolidar mis conocimientos técnicos, para mejorar la planificación, la toma de decisiones y la eficiencia en el desarrollo de nuevos fármacos. Se abordan dos problemas predictivos complementarios relacionados con los ensayos clínicos: (i) la predicción del estado final del ensayo (éxito o fracaso) y (ii) la estimación de la duración total del mismo. Ambos problemas se tratan de forma independiente, utilizando modelos supervisados adaptados a la naturaleza de cada tarea.

1.3. Objetivos del trabajo

Objetivo Principal

El objetivo principal es aplicar técnicas de aprendizaje automático para optimizar el diseño de ensayos clínicos y predecir de manera temprana los resultados de éstos, contribuyendo a reducir riesgos, tiempos y costes en el desarrollo de nuevos medicamentos.

Objetivos secundarios

- Analizar y preparar conjuntos de datos clínicos, garantizando su calidad, consistencia y cumplimiento de buenas prácticas éticas en el manejo de datos de salud.
- Identificar las variables más relevantes de los ensayos clínicos que influyen en los resultados, permitiendo un análisis más profundo y predictivo.
- Explorar distintos modelos supervisados para determinar los más adecuados según métricas de rendimiento como precisión, recall, F1-score y AUC-ROC.
- Proponer recomendaciones sobre cómo integrar los modelos predictivos en la planificación de nuevos ensayos clínicos, apoyando la toma de decisiones basada en datos.

Los objetivos específicos del trabajo son:

- Construir un dataset estructurado a partir de la API de ClinicalTrials.gov.
- Desarrollar un modelo de clasificación para predecir el estado final de un ensayo.
- Desarrollar un modelo de regresión para estimar la duración del ensayo.
- Evaluar e interpretar los modelos mediante métricas y técnicas de explicabilidad.

1.4. Enfoque y método seguido

La estrategia elegida para este proyecto aprovecha datos reales que provienen de bases públicas, garantizando relevancia y reproducibilidad.

Estado del arte y revisión bibliográfica
Revisión de la literatura sobre aprendizaje automático aplicado a ensayos clínicos, identificando trabajos previos, metodologías empleadas y hallazgos relevantes. Esto permite situar el proyecto en el contexto científico actual y establecer buenas prácticas para el análisis de datos clínicos.

Recopilación y preprocesamiento de datos
Se obtendrán datasets públicos de ensayos clínicos, principalmente de `ClinicalTrials.gov`. Los datos serán limpiados y normalizados.

Análisis exploratorio de datos (EDA)
Se explorarán los datasets para identificar variables clave, comprender la distribución de los resultados y detectar posibles inconsistencias o valores atípicos, asegurando que los datos sean adecuados para el modelado.

Preparación de datos
Se realizará una limpieza más profunda, tratamiento de valores ausentes e imputación cuando sea necesario. Se codificarán variables categóricas y se normalizarán las variables numéricas para garantizar consistencia y calidad en los datosy y se seleccionarán las variables relevantes para el análisis predictivo.

Modelización predictiva
Se implementarán técnicas de aprendizaje automático supervisado y no supervisado, incluyendo selección de características, ajuste de hiperparámetros y validación cruzada. Los modelos serán evaluados mediante métricas como precisión, recall, F1-score y AUC-ROC.

Interpretación de resultados
Se identificarán las variables más críticas que influyen en los resultados de los ensayos clínicos y se analizarán los patrones predictivos obtenidos, proporcionando información útil para la planificación de nuevos estudios.

1.5. Planificación del trabajo

1.5.1. Recursos necesarios

Para llevar a cabo el trabajo se requerirán los siguientes recursos:

- **Datos:** Datasets públicos de ensayos clínicos, principalmente de `ClinicalTrials.gov`,

que contengan información sobre fases, duración, tipo de estudio y resultados.

- **Hardware:** Ordenador con al menos 16 GB de RAM y procesador i5/i7 o superior, capaz de manejar grandes volúmenes de datos. Adicionalmente se puede optar por usar herramientas en la nube.
- **Software:** Lenguaje Python con librerías como pandas, scikit-learn, XGBoost, y Jupyter Notebook. Se puede valorar el uso de R.
- **Documentación:** Plantillas LaTeX y herramientas de gestión de referencias para redactar la memoria del TFM.

1.5.2. Planificación temporal de tareas

La planificación del proyecto sigue un flujo orientado a la predicción del tiempo de desarrollo de nuevos fármacos. Se indican tanto la duración estimada de cada fase como las fechas de inicio y entrega.

Fase	Actividad principal	Duración estimada
1	Definición del TFM: enunciado del proyecto y planificación inicial	2 semanas
2	Estado del arte y revisión bibliográfica: identificación de trabajos previos, metodologías y hallazgos relevantes	3 semanas
3	Recopilación y preprocesamiento de datos: obtención de datasets de `ClinicalTrials.gov`, limpieza, normalización y selección de variables	2 semanas
4	Análisis exploratorio de datos (EDA) y preparación para modelado: identificación de variables clave, codificación y tratamiento de valores ausentes	2 semanas
5	Desarrollo de modelos de ML: implementación de modelos supervisados, ajuste de hiperparámetros y validación cruzada	2 semanas
6	Evaluación y comparación de modelos e interpretación de resultados: métricas, identificación de variables críticas y patrones predictivos	1 semana
7	Redacción de memoria: entrega final	1 semana
8	Presentación audiovisual y entrega al tribunal, preparación para defensa	2 semanas
9	Defensa pública del trabajo	3 semanas

1.5.3. Riesgos y planes de mitigación

- **Riesgo:** Datos incompletos o inconsistentes sobre duración de fases de ensayos clínicos. **Mitigación:** Seleccionar múltiples datasets públicos, de diferentes fuentes, realizar limpieza exhaustiva y documentar transformaciones.
- **Riesgo:** Modelos de ML con bajo rendimiento en predicción temporal. **Mitigación:** Probar distintos algoritmos, ajustar hiperparámetros y aplicar técnicas de feature engineering orientadas a duración. Revisar bibliografía existente para definir mejor los modelos.
- **Riesgo:** Retrasos en la planificación. **Mitigación:** Priorizar tareas críticas, seguimiento semanal y recalcular la planificación.

1.5.4. Duración total estimada

La duración total estimada del proyecto es de 14 semanas (aproximadamente 3.5 meses), abarcando desde la revisión bibliográfica hasta la entrega final, incluyendo todas las fases de análisis de datos, modelización y documentación.

Capítulo 2

Estado del arte

2.1. Introducción

En este capítulo se presenta una revisión de cómo otros investigadores han abordado la aplicación de técnicas de ML (*machine learning*) en el ámbito de los ensayos clínicos.

La aplicación de ML en el ámbito clínico se ha consolidado en los últimos años como una herramienta clave. Se ha usado para optimizar procesos, reducir costes y mejorar la eficiencia del desarrollo farmacéutico. Su integración permite analizar grandes volúmenes de datos clínicos, identificar patrones complejos y predecir resultados que tradicionalmente requerían largos periodos de observación o coste.

Para realizar una búsqueda exhaustiva de la bibliografía existente, se ha utilizado la herramienta *Google Scholar*, la cual permite localizar de manera sencilla y eficiente publicaciones académicas.

Para acotar los resultados, se han empleado palabras clave (*keywords*) como por ejemplo *çlinical trial prediction", "predictive modelling trial outcomes.º çlinical phase transition"*. Además, para ampliar el análisis se han revisado otros artículos relacionados que aparecen en la sección .Artículos relacionados".

Se han analizado las fases típicas de los ensayos clínicos y las soluciones propuestas mediante aprendizaje automático. El objetivo ha sido identificar las tendencias y los enfoques predominantes en la literatura existente, y las principales brechas de investigación que fundamentan este trabajo.

2.2. Ensayos clínicos: fases, estructura y retos

Los ensayos clínicos son estudios diseñados para descubrir o verificar los efectos de uno o más medicamentos en investigación. La Agencia Europea de Medicamentos (EMA) [4] se basa en los resultados de estos ensayos, realizados por las compañías farmacéuticas, para emitir sus dictámenes sobre la autorización de medicamentos.

A continuación se describen las diferentes fases de un ensayo clínico[2]:

1. **Fase I.** En esta fase se administra un nuevo medicamento a seres humanos por primera vez, generalmente a voluntarios sanos. Se analiza cómo el organismo procesa el medicamento, sus principales efectos y sus principales efectos secundarios.
2. **Fase II.** Se realiza después de los estudios de fase I para evaluar los efectos de un medicamento en una condición particular y determinar sus efectos secundarios comunes a corto plazo.
3. **Fase III.** Se suele realizar en un gran grupo de pacientes para confirmar la eficacia y seguridad de un medicamento, y así poder evaluar sus beneficios y riesgos.
4. **Fase IV.** Se realiza después de la autorización de un medicamento, con el objetivo de vigilar efectos adversos a largo plazo y evaluar la efectividad en la práctica clínica habitual.

La siguiente tabla (2.1) muestra las características principales de las cuatro fases (I-IV) típicas de un ensayo clínico (elaboración propia a partir de la FDA (*U.S. Food and Drug Administration* [3]) y la EMA (*European Medicines Agency* [4]):

Cuadro 2.1: *Fases de los ensayos clínicos y sus características principales*

Fase	Participantes	Duración del estudio	Objetivo principal	Comentario
Fase I	De 20 a 100 voluntarios sanos o personas con la enfermedad/afección.	Varios meses	Evaluar la seguridad y determinar la dosificación adecuada.	Aproximadamente el 70 % de los fármacos pasan a la siguiente fase.
Fase II	Hasta varios cientos de personas con la enfermedad/afección.	De varios meses a 2 años	Analizar la eficacia preliminar y el perfil de seguridad (efectos secundarios).	Aproximadamente el 33 % de los fármacos pasan a la siguiente fase.
Fase III	De 300 a 3 000 voluntarios con la enfermedad o afección.	De 1 a 4 años	Confirmar la eficacia y monitorizar reacciones adversas.	Aproximadamente entre el 25 % y el 30 % de los fármacos pasan a la siguiente fase.
Fase IV	Varios miles de voluntarios con la enfermedad/afección.	Indefinida	Detectar efectos adversos raros y evaluar la efectividad real.	Fase posterior a la comercialización (*post-marketing*).

2.3. Aplicaciones del machine learning en ensayos clínicos

En esta sección se presenta una revisión de los principales estudios que han aplicado técnicas de aprendizaje automático en el ámbito de los ensayos clínicos:

Cuadro 2.2: Estudios relevantes sobre predicción de resultados en ensayos clínicos mediante aprendizaje automático

Paper	Técnica utilizada	Fase centrada
Improving clinical trial design using interpretable ML based prediction of early trial termination [5]	Logistic regression, Random Forest y xgBoost	I-II
Predicting clinical trial duration via statistical and ML models [6]	Decision Tree, Neural Networks y SSVM	I-III
SPOT: Sequential Predictive Modeling of Clinical Trial Outcome with Meta-Learning [7]	Meta-learning, Redes neuronales	I-III
Key indicators of phase transition for clinical trials through ML [8]	Random Forest	II-III
Prediction of Clinical Trials Outcomes Based on Target Choice and Clinical Trial Design with Multi-Modal AI [9]	Deep Learning + ML	I-III
Predicting the Outcome of Phase III Trials using Phase II Data [10]	Modelado y simulación clínica, regresión	III
A scoping review of AI applications in clinical trial risk assessment [11]	Revisión sistemática, ML, NLP, Deep Learning	Todas
The role of machine learning in clinical research - BioMed Central [12]	Revisión general, ML supervisado y no supervisado	Todas
A Survey of Artificial Intelligence Methods for Clinical Trial Outcome Prediction [13]	Métodos AI, ML supervisado y Deep Learning	Todas
Predicting Clinical Trial Completion and Success Using ML + NLP [14]	ML + NLP	I-III

Las fuentes de datos combinan tanto registros públicos como privados, con ClinicalTrials.gov como principal referencia, la cual proporciona registros históricos de los ensayos clínicos de fases I a III. Estos datos se complementan con datos propios de la industria farmacéutica ([7],[10],[13],[14]), o con DrugBank y PubMed ([9]).

2.4. Modelos y enfoques existentes

En esta sección se describen los principales modelos de aprendizaje automático aplicados en ensayos clínicos.

Los modelos supervisados clásicos, como regresión logística permiten para predecir probabilidad de éxito o fracaso de un ensayo; útiles en fases I-II. Los árboles de decisión y Random Forest se usan para identificar variables clave y relaciones no lineales de forma interpretable. Por último las Support Vector Machines normalmente se usan para clasificación de resultados de ensayos. En fases avanzadas (III y IV), se emplean redes neuronales profundas (DNN) y Graph Neural Networks (GNN) para capturar relaciones complejas entre datos clínicos, moleculares y regulatorias, mejorando así la predicción de resultados y la duración de los ensayos, eficacia terapéutica, la tasa de abandono o el éxito de transición entre fases.

Uno de los principales cuellos de botella de los ensayos clínicos es el reclutamiento adecuado de pacientes, que puede representar hasta el 30-40 % del coste total del estudio [15]. El procesamiento del lenguaje natural (NLP) se utilizan para analizar historiales médicos electrónicos (EHRs) y detectar automáticamente candidatos potenciales basándose en patrones de texto clínico. Posteriormente, modelos de clasificación supervisada - como Random Forest o Gradient Boosting- logran identificar individuos elegibles con alta precisión, reduciendo los tiempos de reclutamiento (SPOT).

La monitorización de eventos adversos durante el desarrollo clínico es un componente esencial de la seguridad farmacológica. El NLP se ha aplicado para extraer automáticamente reportes de seguridad de bases de datos regulatorias, foros médicos y literatura científica. Estas técnicas contribuyen a la detección temprana de señales de seguridad y al cumplimiento de normativas regulatorias como las de la FDA o la EMA.

Cuadro 2.3: Principales áreas de aplicación del *machine learning* en ensayos clínicos

Subárea	**Ejemplo de aplicación**	**Tipo de técnica ML**
Diseño y planificación	Selección de criterios de inclusión/exclusión	Regresión logística, Árboles de decisión
Reclutamiento de pacientes	Identificación de candidatos mediante EHRs	Random Forest, Gradient Boosting, NLP
Predicción de resultados	Éxito de transición de fase y tasa de abandono	XGBoost, Redes neuronales profundas
Detección de eventos adversos	Extracción automática de reportes clínicos	NLP

2.5. Retos actuales y líneas futuras

Los principales retos se relacionan con la heterogeneidad y la baja calidad de los datos disponibles, la interpretabilidad de los modelos, la validación con datos del mundo real (Real World Data, RWD) así como con la falta de estandarización en las variables, lo que provoca sesgos que influyen en la fiabilidad de las predicciones.

La integración de fuentes heterogéneas - clínicas, moleculares y textuales- y la aplicación de técnicas de NLP han favorecido el analisis de informes de efectos adversos, mientras que la tendencia hacia modelos explicables busca aumentar la confianza y la aplicabilidad regulatoria.La tendencia actual apunta hacia modelos híbridos explicables, capaces de integrar información multifuente, capturar relaciones no lineales.

Esta estado justifica el presente TFM, orientado a desarrollar y evaluar modelos predictivos reproducibles y explicables que permitan identificar los factores asociados al éxito o fracaso de los ensayos clínicos.

Capítulo 3

Materiales y Métodos

3.1. Diseño del sistema

Antes de la fase de implementación se realizó una etapa de diseño del sistema, con el objetivo de definir claramente el alcance del trabajo, los requisitos y la estructura general de la solución. Esta fase permitió identificar los principales bloques funcionales del proyecto y reducir la complejidad durante la implementación.

A partir de este diseño se definió un flujo de trabajo compuesto por las siguientes etapas: (i) descarga y almacenamiento de datos, (ii) limpieza y preprocesamiento, (iii) ingeniería de características, (iv) modelado predictivo y (v) evaluación e interpretabilidad.

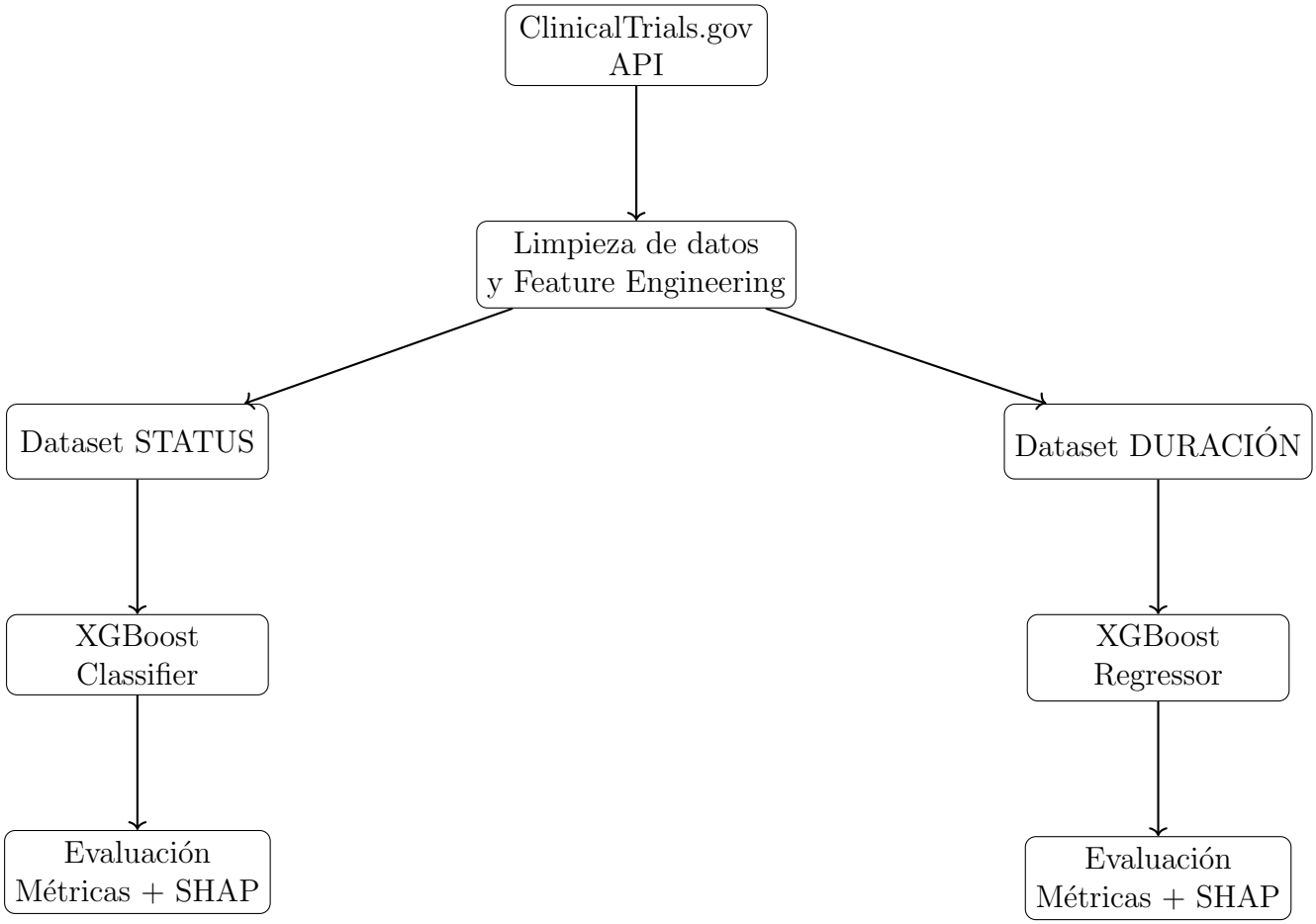

Figura 3.1: Diagrama de bloques del sistema propuesto para la predicción del estado y la duración de ensayos clínicos.

3.2. Fuente de datos

Los datos utilizados en este trabajo proceden del registro público *ClinicalTrials.gov*, mantenido por la *U.S. National Library of Medicine*. Esta plataforma proporciona información estructurada sobre ensayos clínicos a nivel mundial, incluyendo variables relacionadas con el diseño del estudio, estado, fase clínica, patrocinador, fechas clave y características demográficas.

La descarga de los datos se realizó mediante la API oficial de ClinicalTrials.gov, lo que garantiza la reproducibilidad del proceso. Se seleccionaron principalmente ensayos correspondientes a las fases I–IV, excluyendo estudios con información incompleta o inconsistencias graves en variables temporales.

La base de datos ClinicalTrials.gov resulta especialmente adecuada para este estudio debido a su carácter público, su cobertura internacional y la disponibilidad de información estructurada sobre el diseño y resultado de los ensayos clínicos.

3.3. Preprocesamiento y limpieza de datos

El preprocesamiento de los datos se llevó a cabo siguiendo buenas prácticas de ingeniería de datos y aprendizaje automático. Entre las principales tareas realizadas se incluyen:

- Normalización de los nombres de las variables.
- Conversión explícita de variables booleanas a formato numérico.
- Tratamiento de valores ausentes y eliminación de registros no informativos.
- Eliminación de variables de alta cardinalidad no directamente utilizables en esta fase.
- Transformación de variables temporales en métricas numéricas (años, meses y duraciones en días).

Las variables categóricas se codificaron mediante variables binarias (dummy variables) para facilitar su uso en modelos basados en árboles. Asimismo, se descartaron variables de texto libre para evitar introducir ruido en esta fase del trabajo.

Como resultado de este proceso, se construyeron dos conjuntos de datos finales independientes: uno orientado a la predicción del estado del ensayo y otro a la estimación de su duración.

3.4. Definición de los problemas de modelado

3.4.1. Predicción del estado del ensayo

El primer problema se formuló como una tarea de clasificación binaria. La variable objetivo toma el valor 1 para ensayos con estado *COMPLETED* y el valor 0 para aquellos con estado *TERMINATED*, *WITHDRAWN* o *SUSPENDED*.

El conjunto de datos final para este problema consta de 28 956 observaciones y 45 variables explicativas.

3.4.2. Predicción de la duración del ensayo

El segundo problema se formuló como una tarea de regresión supervisada, donde la variable objetivo es la duración primaria del ensayo en días (*DurationPrimaryDays*). Este conjunto de datos contiene 40 935 observaciones y 45 variables explicativas.

3.5. Modelos empleados

Para ambos problemas se utilizó el algoritmo XGBoost, debido a su buen rendimiento en problemas tabulares, su capacidad para modelar relaciones no lineales y su compatibilidad con técnicas de interpretabilidad.

- Clasificación: `XGBClassifier`
- Regresión: `XGBRegressor`

Los hiperparámetros se ajustaron mediante búsqueda aleatoria y validación cruzada en etapas previas del trabajo.

Tras evaluar distintos modelos, XGBoost fue seleccionado como modelo final para la tarea de clasificación debido a su buen rendimiento en términos de AUC y a su capacidad para manejar relaciones no lineales y datasets desbalanceados.

El problema de duración se abordó como una tarea de regresión, donde el objetivo era estimar el número de días hasta la finalización primaria del ensayo clínico.

3.6. Evaluación y métricas

En el problema de clasificación se emplearon las métricas *Accuracy*, *Precision*, *Recall*, *F1-score* y *AUC-ROC*. Además, se analizaron distintos umbrales de decisión con el objetivo de estudiar el compromiso entre sensibilidad y especificidad.

Para el problema de regresión se utilizaron las métricas *RMSE*, *MAE* y el coeficiente de determinación R^2, complementadas con análisis visual de residuos.

3.7. Interpretabilidad

Con el objetivo de garantizar la explicabilidad de los modelos, se aplicaron técnicas basadas en SHAP (*SHapley Additive exPlanations*). Se realizaron análisis globales y locales, permitiendo identificar qué variables influyen más en las predicciones y en qué dirección.

El análisis de interpretabilidad mediante valores SHAP permite comprender qué variables influyen en las predicciones del modelo, aumentando la transparencia y la confianza en los resultados obtenidos.

Las variables más relevantes identificadas (como el número de participantes o la fase del ensayo) resultan coherentes desde un punto de vista clínico y regulatorio.

Capítulo 4

Resultados

4.1. Resultados del modelo de clasificación

El modelo de clasificación basado en XGBoost alcanzó un valor de AUC de 0.842, lo que indica una elevada capacidad discriminativa entre ensayos exitosos y fallidos.

El modelo alcanza un valor de AUC de 0.84, lo que indica una buena capacidad de discriminación entre ensayos exitosos y fallidos. Sin embargo, se observa un compromiso entre la sensibilidad de la clase minoritaria (fracaso del ensayo) y la precisión global, especialmente al modificar el umbral de decisión.

La Figura 4.1 muestra la curva ROC obtenida para el modelo final.

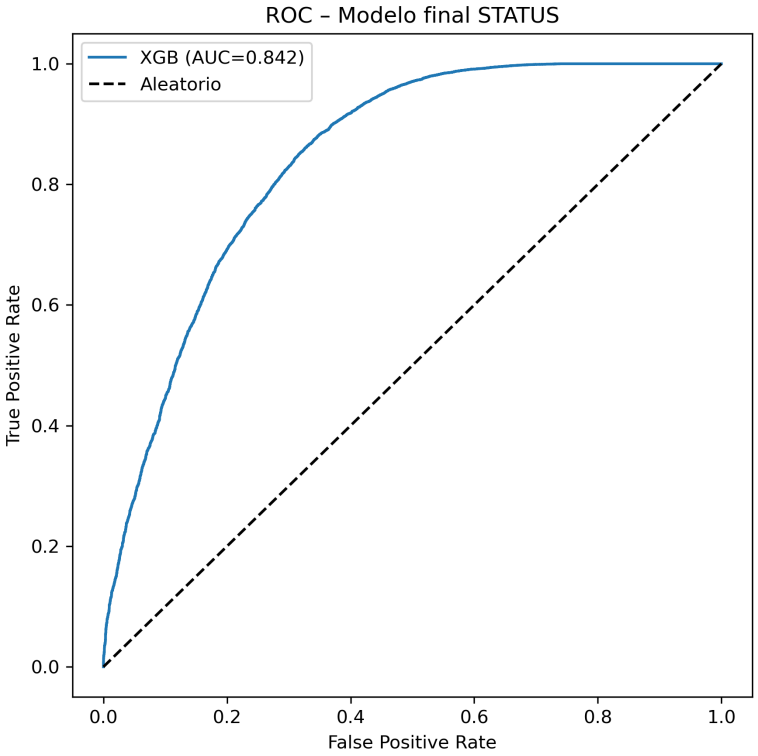

Figura 4.1: Curva ROC del modelo XGBoost para la predicción del estado del ensayo.

La Figura 4.2 presenta la matriz de confusión para un umbral de decisión de 0.5.

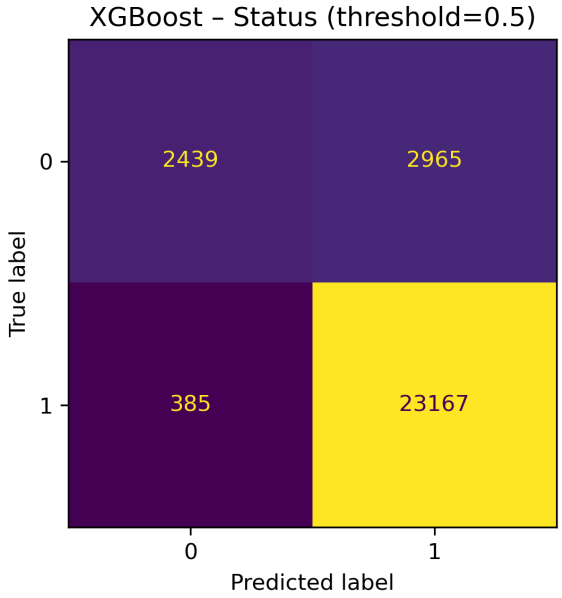

Figura 4.2: Matriz de confusión del modelo de clasificación (umbral = 0.5).

4.2. Interpretabilidad del modelo de clasificación

El análisis de interpretabilidad global mediante SHAP revela las variables con mayor impacto en la predicción del éxito del ensayo. La Figura 4.3 muestra el gráfico *beeswarm*, donde se observa la contribución de cada variable al resultado del modelo.

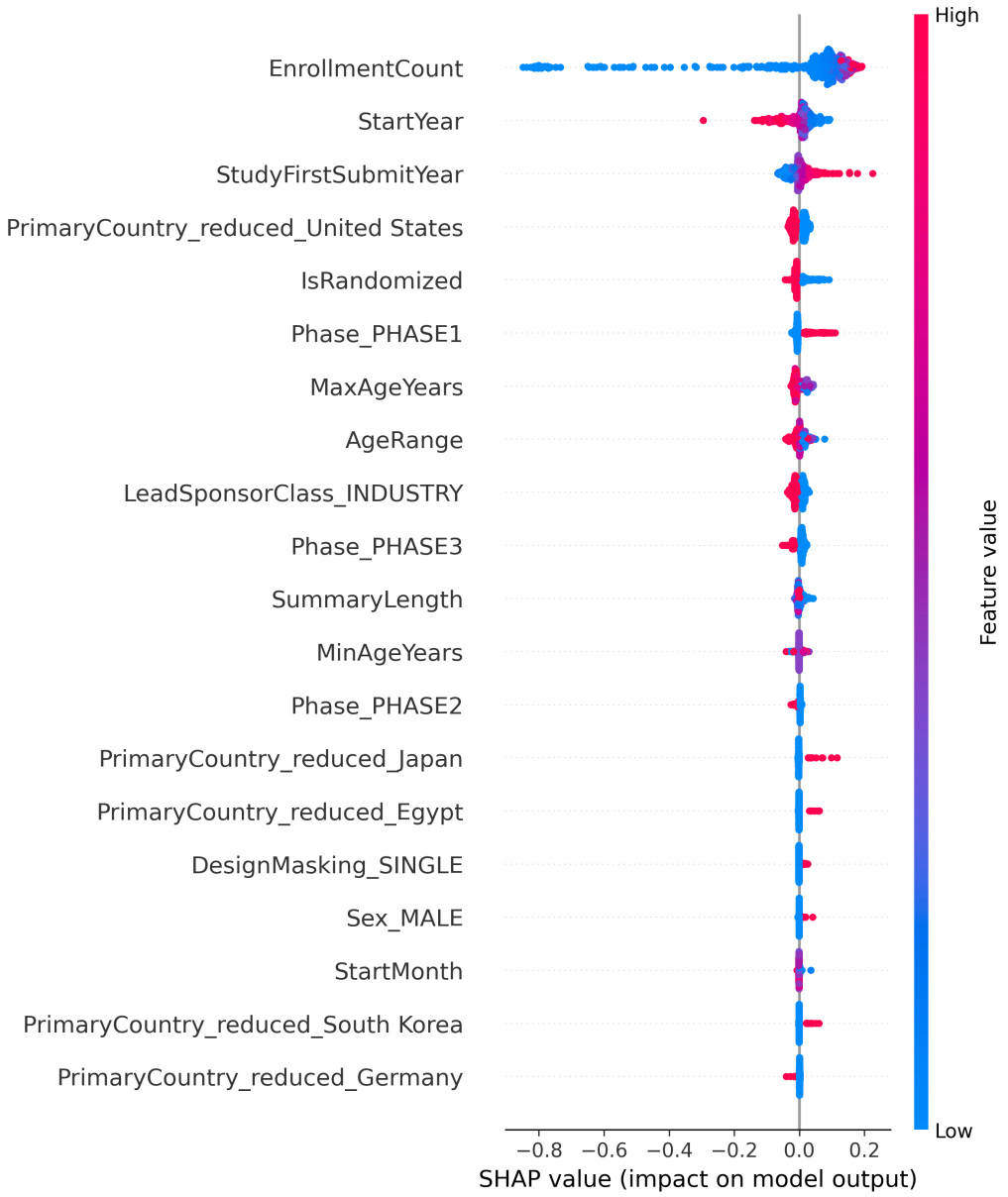

Figura 4.3: Importancia global de variables mediante SHAP (clasificación).

La Figura 4.4 ilustra un ejemplo de interpretabilidad local mediante un gráfico *waterfall*, correspondiente a un ensayo clasificado como fallido.

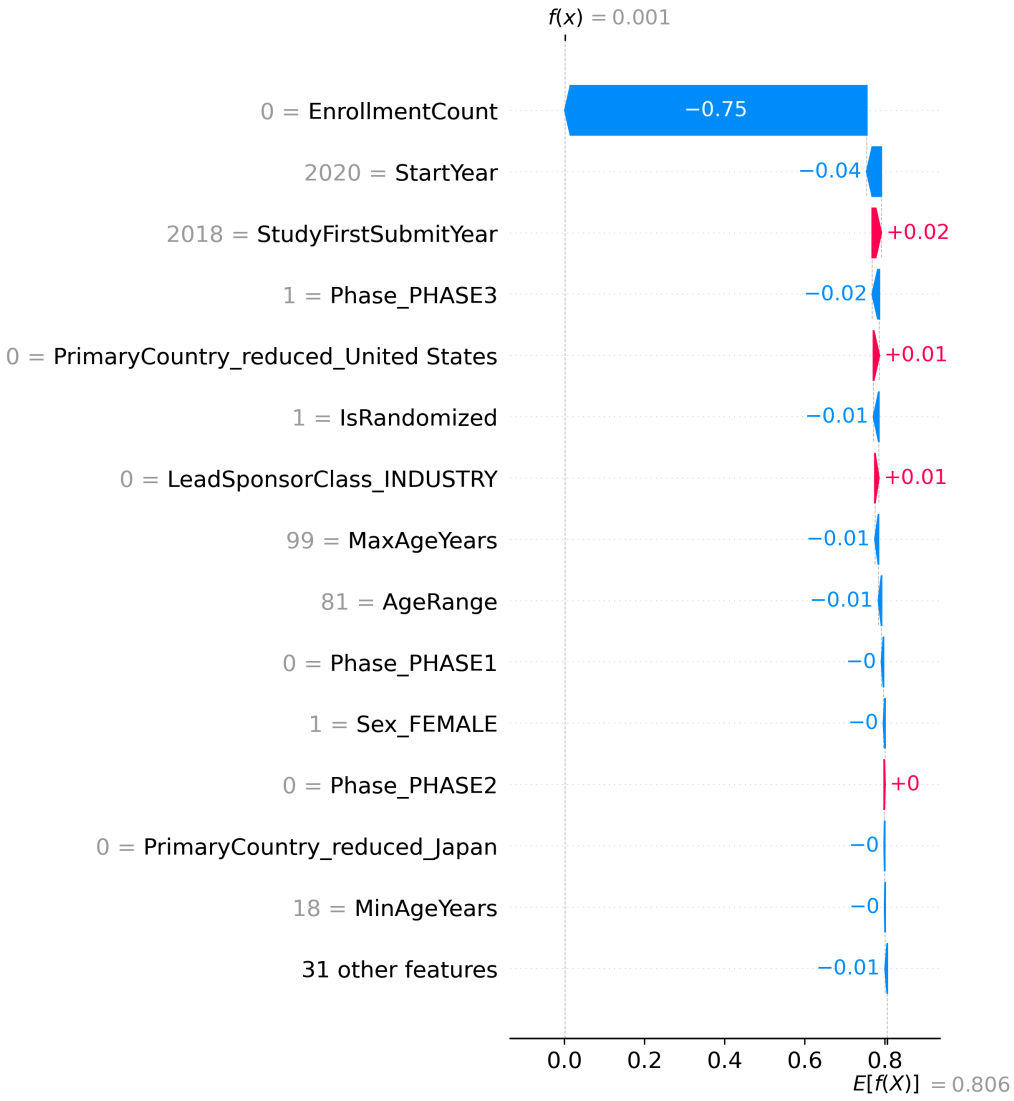

Figura 4.4: Interpretabilidad local mediante SHAP para un ensayo clasificado como fracaso.

4.3. Resultados del modelo de regresión

El modelo de regresión XGBRegressor obtuvo un RMSE aproximado de 658 días, un MAE de 452 días y un coeficiente de determinación R^2 de 0.34.

En el caso de la regresión, el modelo obtiene un R^2 de aproximadamente 0.34, lo que sugiere que una parte relevante de la variabilidad en la duración del ensayo puede explicarse a partir de las variables disponibles, aunque existen factores externos no capturados por el modelo.

La Figura 4.5 muestra la comparación entre los valores reales y los valores predichos por el modelo.

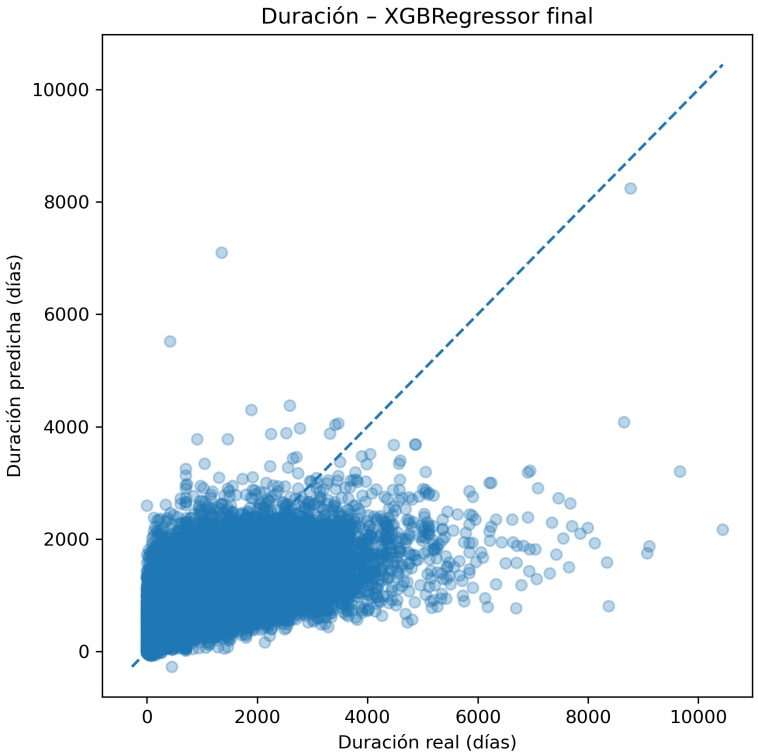

Figura 4.5: Comparación entre duración real y duración predicha del ensayo.

4.4. Interpretabilidad del modelo de duración

El análisis SHAP aplicado al modelo de regresión permitió identificar los factores que más influyen en la duración de los ensayos clínicos. La Figura 4.6 muestra la importancia global de las variables.

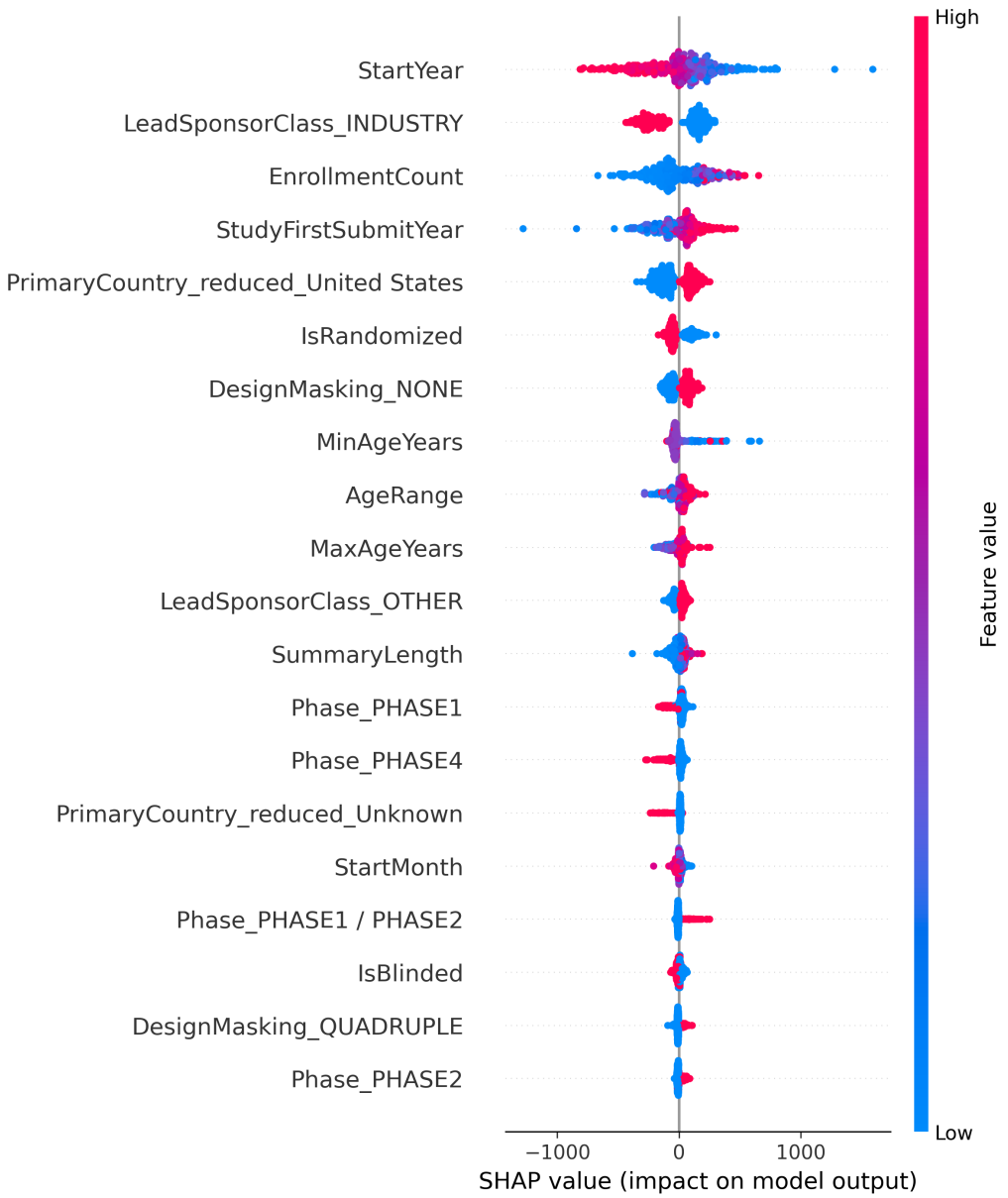

Figura 4.6: Importancia global de variables mediante SHAP para el modelo de duración.

Capítulo 5

Discusión

Entre las principales limitaciones del estudio se encuentra la ausencia de información clínica detallada de los pacientes y la posible presencia de sesgos en los datos históricos disponibles.

En este trabajo se ha abordado la predicción temprana del resultado y la duración de ensayos clínicos utilizando técnicas de *Machine Learning* aplicadas a datos públicos procedentes de la base de datos *ClinicalTrials.gov*. Los resultados obtenidos permiten extraer varias conclusiones relevantes, así como identificar limitaciones y líneas de mejora futuras.

En primer lugar, en el problema de clasificación del estado final del ensayo clínico (éxito frente a fracaso), el modelo basado en *XGBoost* ha alcanzado un valor de AUC cercano a 0.84, lo que indica una capacidad discriminativa sólida teniendo en cuenta el desbalanceo de clases presente en el conjunto de datos. El análisis de diferentes umbrales de decisión ha mostrado que existe un compromiso claro entre la sensibilidad para detectar ensayos fallidos y la precisión global del modelo. En particular, el uso de umbrales más conservadores permite mejorar el *recall* de la clase minoritaria (ensayos no completados), lo cual puede ser especialmente relevante en contextos de planificación temprana, donde el coste de no detectar un posible fracaso puede ser elevado.

Por otro lado, el análisis de interpretabilidad mediante valores SHAP ha permitido identificar de forma consistente las variables más influyentes en la predicción del estado del ensayo. Entre ellas destacan el tamaño de la muestra (*EnrollmentCount*), la fase clínica, el tipo de patrocinador y determinadas características demográficas de la población objetivo. Estos resultados son coherentes con el conocimiento experto del dominio clínico, lo que refuerza la validez del modelo y su potencial utilidad práctica.

En el problema de regresión para la predicción de la duración del ensayo clínico, el modelo *XGBRegressor* ha obtenido un coeficiente de determinación R^2 en torno a 0.34. Aunque este valor puede considerarse moderado desde una perspectiva puramente estadística, resulta razonable dada la alta variabilidad inherente a la duración de los ensayos clínicos y la ausencia de variables clave no disponibles en la base de datos pública, como factores regulatorios, decisiones

estratégicas internas o eventos imprevistos durante la ejecución del estudio. En este sentido, el modelo proporciona una estimación aproximada de la duración esperada, más adecuada para análisis exploratorios y comparativos que para predicciones exactas a nivel individual.

El análisis SHAP aplicado al modelo de duración ha mostrado que variables relacionadas con la fase del ensayo, el tipo de patrocinador y el tamaño de la muestra vuelven a desempeñar un papel relevante, lo que sugiere una coherencia estructural entre ambos problemas abordados en el trabajo. Además, los gráficos de residuos indican la presencia de una dispersión considerable, lo que refleja la complejidad del fenómeno modelado y confirma que una parte significativa de la variabilidad no puede ser explicada únicamente a partir de los datos disponibles.

Entre las principales limitaciones del trabajo destaca la dependencia de una única fuente de datos pública, lo que implica posibles sesgos de reporte, valores faltantes y heterogeneidad en la calidad de la información. Asimismo, el enfoque se ha centrado en variables estructuradas de tipo tabular, dejando fuera información textual potencialmente relevante, como descripciones detalladas del protocolo o criterios de inclusión y exclusión, que podrían ser explotadas mediante técnicas de *Natural Language Processing* en trabajos futuros.

Como líneas de mejora, se propone la incorporación de variables adicionales procedentes de fuentes externas, el uso de modelos multimodales que combinen información estructurada y no estructurada, así como la evaluación del enfoque en subconjuntos específicos de ensayos clínicos (por ejemplo, por área terapéutica o fase). En conjunto, los resultados obtenidos demuestran que el uso de técnicas de *Machine Learning* interpretables sobre datos públicos puede aportar valor en el análisis temprano de ensayos clínicos, sentando las bases para desarrollos más avanzados en trabajos posteriores.

Capítulo 6

Bibliografía

1. National Library of Medicine (EE. UU.). ClinicalTrials.gov [Internet]. Disponible en: https://clinicaltrials.gov/about-site/about-ctg [consultado el 11 de octubre de 2025].

2. U.S. Food and Drug Administration. Clinical Trials and Human Subject Protection [Internet]. Disponible en: https://www.fda.gov/science-research/science-and-research-special-topics/clinical-trials-and-human-subject-protection [consultado el 11 de octubre de 2025].

3. U.S. Food and Drug Administration. The drug development process [Internet]. Disponible en: https://www.fda.gov/patients/drug-development-process [consultado el 1 de noviembre de 2025].

4. European Medicines Agency. Clinical trials: Regulation and guidance [Internet]. Disponible en: https://www.ema.europa.eu/en/human-regulatory/research-development/clinical-trials [consultado el 1 de noviembre de 2025].

5. Kavalci, E., Hartson, A. Improving clinical trial design using interpretable machine learning based prediction of early trial termination [Internet]. Scientific Reports. Disponible en: https://doi.org/10.1038/s41598-023-27416-7 [consultado el 2 de noviembre de 2025].

6. Cho, J., Xu, Q., Wong, C. Predicting clinical trial duration via statistical and machine learning models [Internet]. Computers in Biology and Medicine. Disponible en: https://doi.org/10.1016/j.conctc.2025.101473 [consultado el 2 de noviembre de 2025].

7. Wang, Z., Xiao, C., Sun, J. SPOT: Sequential Predictive Modeling of Clinical Trial Outcome with Meta-Learning [Internet]. ACM Digital Library. Disponible en: https://doi.org/10.1145/3584371.3613001 [consultado el 2 de noviembre de 2025].

8. Feijoo, F., Palopoli, M., et al,. Key indicators of phase transition for clinical trials through machine learning [Internet]. Drug Discovery Today. Disponible en: https://doi.org/10.1016/j.drudis.2019.12.014 [consultado el 2 de noviembre de 2025].

9. Aliper, A., Kudrin, R et al. Prediction of Clinical Trials Outcomes Based on Target Choice and Clinical Trial Design with Multi-Modal Artificial Intelligence [Internet]. CPT: Pharmacometrics & Systems Pharmacology. Disponible en: https://doi.org/10.1002/cpt.3008 [consultado el 2 de noviembre de 2025].

10. De Ridder, F. Predicting the Outcome of Phase III Trials using Phase II Data: A Case Study of Clinical Trial Simulation in Late Stage Drug Development [Internet]. Naunyn-Schmiedeberg's Archives of Pharmacology. Disponible en: https://doi.org/10.1111/j.1742-7843.2005.pto960314.x [consultado el 2 de noviembre de 2025].

11. Teodoro, D., Naderi, N. et al. A scoping review of artificial intelligence applications in clinical trial risk assessment [Internet]. npj Digital Medicine. Disponible en: https://doi.org/10.1038/s41746-025-01886-7 [consultado el 2 de noviembre de 2025].

12. Weissler, E.H., Naumann, T., et al. The role of machine learning in clinical research: transforming the future of evidence generation - BioMed Central [Internet]. Trials. Disponible en: https://doi.org/10.1186/s13063-021-05489-x [consultado el 2 de noviembre de 2025].

13. Qian, L., Lu, X., et al. A Survey of Artificial Intelligence Methods for Clinical Trial Outcome Prediction [Internet]. ChemRxiv. Disponible en: https://doi.org/10.26434/chemrxiv-2024-08t4w [consultado el 2 de noviembre de 2025].

14. Jiazheng Li. Predicting Clinical Trial Completion and Success Using Machine Learning and NLP [Internet]. University of Chicago. Disponible en: https://knowledge.uchicago.edu/record/15346/files/thesis.pdf [consultado el 2 de noviembre de 2025].

15. Lluch, J. Trial Phase Costing Benchmarks: How Much Should Phase II Really Cost in 2025? [Internet]. Abacum. Disponible en: https://www.abacum.ai/blog/trial-phase-costing-benchmarks [consultado el 2 de noviembre de 2025].