



UNIVERSITAT  
ROVIRA I VIRGILI

**FARMACODB: AUTOMATIZACIÓN DE EXTRACCIÓN Y ANÁLISIS DE  
DATOS FARMACOCINÉTICOS**

**Pablo Sierra Fernández**

**TRABAJO FINAL DE GRADO BIOTECNOLOGÍA**

**Tutor académico:** Santiago García Vallvé, Bioquímica i Biotecnologia,  
[santi.garcia-vallve@urv.cat](mailto:santi.garcia-vallve@urv.cat)

**En cooperació amb:** Anaxomics Biotech SL

**Supervisor:** Pedro M. Filipe, Anaxomics, [pedro.filipe@anaxomics.com](mailto:pedro.filipe@anaxomics.com)

Fecha de convocatoria; Junio 2024

Yo, *Pablo Sierra Fernández*, con DNI 54191004-Z, soy conocedor de la guía de prevención del plagio de la URV *Prevenió, detecció i tractament del plagi en la docència: guia per a estudiants* (aprobada en julio de 2017) (<https://www.crai.urv.cat/ca/serveis/suport-aprenentatge/plagi/>) y afirmo que este TFG no constituye ninguna de las conductas consideradas como plagio por la URV.


Tarragona, 5 de marzo de 2024



<b>1. Datos del centro.....</b>	<b>4</b>
<b>2. Resumen.....</b>	<b>5</b>
<b>3. Introducción.....</b>	<b>6</b>
3.1 Principales Parámetros Farmacocinéticos (PK).....	6
3.2 Inteligencia Artificial.....	8
3.3 Modelos de lenguaje.....	8
3.4 Problemática.....	9
<b>4. Objetivos.....</b>	<b>10</b>
<b>5. Metodología.....</b>	<b>11</b>
5.1 Creación de la base de datos.....	11
5.1.1 Extracción de datos PK de PKPDAI y DrugBank.....	12
5.1.2 Minería de texto para extracción de "especie" y "enfermedad".....	13
5.1.3 Extracción de Descriptores Moleculares.....	14
5.1.4 Limpieza de Datos.....	16
5.1.5 Normalización de Unidades y Valores.....	18
5.1.6 Integración de Parámetros Descriptivos.....	19
5.2 Desarrollo del Modelo Predictivo.....	20
5.2.1 Limpieza de datos.....	20
5.2.2 Normalización y Estandarización.....	21
5.2.3 Modelos Utilizados.....	22
5.2.4 Implementación de Librerías de Optimización.....	23
5.2.5 Selección de Características y Modelos.....	23
<b>6. Resultados y Discusión.....</b>	<b>25</b>
6.1 Creación de base de datos.....	25
6.1.1 Minería de Texto y Modelos de Lenguaje.....	26
6.1.2 Evaluación de Modelos y Precisión.....	27
6.1.3 Cálculo de Precisión (Accuracy).....	27
6.1.4 Descriptores Moleculares.....	27
6.1.5 Normalización de Nombres de Parámetros.....	28
6.1.6 Grupos de Sinónimos.....	29
6.1.7 Variabilidad de Unidades.....	30
6.2 Desarrollo del Modelo Predictivo.....	32
6.2.1 Limpieza de Datos.....	33
6.2.2 Modelos Utilizados.....	35
6.2.3 Selección de Características.....	35
6.2.4 Modelos Finalistas.....	37
<b>7. Conclusiones.....</b>	<b>40</b>
7.1 Creación de la base de datos.....	41
7.2 Creación del modelo predictivo.....	42
<b>8. Bibliografía.....</b>	<b>43</b>
<b>9. Autoevaluación.....</b>	<b>44</b>
<b>10. Anexos.....</b>	<b>45</b>

# 1. Datos del centro

**Anaxomics Biotech SL** (<https://www.anaxomics.com/>) es una empresa biotecnológica innovadora, especializada en la aplicación de la Bioinformática y la Biología de Sistemas para superar desafíos en el descubrimiento y desarrollo de fármacos. Utilizando tecnologías de vanguardia basadas en la Biología de Sistemas, Anaxomics se esfuerza por ofrecer soluciones *in silico* para ensayos clínicos y más allá, contribuyendo al avance de la medicina personalizada y la investigación biomédica.

**Información General**

- **Nombre:** Anaxomics Biotech SL
- **Fundador y CSO:** José Manuel Mas
- **Ubicación:** Barcelona, España (sede principal); Valais, Suiza (ubicación adicional)
- **Áreas de Especialización:** Bioinformática, Biología de Sistemas
- **Tecnología Propietaria:** Therapeutic Performance Mapping System (TPMS)
- **Enfoque:** Identificación de nuevas dianas terapéuticas, diseño de fármacos
- **Proyectos Destacados:** Participación en el proyecto LEGACy para la prevención y tratamiento del cáncer gástrico; colaboración en investigaciones sobre ELA con la Universitat Autònoma de Barcelona
- **Publicaciones y Patentes:** Contribución a más de 85 publicaciones y patentes
- **CIF:** B64630569
- **Dirección:** Calle Diputació, 237 - P. 1 Pta. 1, 08007 Barcelona, España
- **Teléfono:** +34 934 516 717

(*Anaxomics Biotech SL - In Silico Clinical Trials And Beyond, s. f.*)

## 2. Resumen

---

La farmacología individualizada se enfrenta a un desafío debido a la falta de bases de datos completas y estandarizadas de parámetros farmacocinéticos. La dispersión de datos farmacocinéticos en la literatura científica, así como en estudios preclínicos y clínicos, complica la recopilación y el acceso a esta información, dificultando su integración y comparación. En este contexto, este trabajo se propuso desarrollar una base de datos integral y estandarizada de parámetros farmacocinéticos, así como crear un modelo predictivo para dichos parámetros, centrándose en la vida media como parámetro principal. Para lograr este objetivo, se desarrolló una base de datos que actualmente contiene información detallada de 1,130 fármacos, representados por 100,359 registros de parámetros farmacocinéticos, utilizando técnicas avanzadas de minería de texto y modelos de lenguaje para la extracción automatizada de información clave. Además, se implementaron estrategias de normalización y estandarización de datos para mejorar la coherencia en la presentación de datos y facilitar su integración y comparación. Paralelamente, se optimizaron las técnicas de minería de texto y modelos de lenguaje para una extracción automatizada más precisa de información clave.

Con respecto al modelo predictivo, se desarrolló un modelo específico para la farmacocinética humana, enfocado en predecir la vida media de los medicamentos a partir de descriptores moleculares. Se aplicaron diversas técnicas de preprocesamiento y selección de características, y se probaron varios modelos, como Regresión Lineal Múltiple y Redes Neuronales Artificiales. Los modelos *ANN 3* y *MLJAR* destacaron como opciones robustas, mostrando un equilibrio entre precisión y capacidad de generalización.

En resumen, este trabajo contribuye al avance de la farmacología individualizada al proporcionar una base de datos integral y estandarizada de parámetros farmacocinéticos y un modelo predictivo robusto para la vida media de los medicamentos. Estos resultados sientan las bases para análisis más profundos en el campo de la farmacocinética y promueven la transparencia y el acceso abierto a los métodos y datos utilizados en la investigación.

**Palabras clave:** farmacología individualizada, PBPK, bases de datos farmacocinéticas, minería de texto, modelos de lenguaje, vida media, modelo predictivo, normalización de datos, estandarización, selección de características, redes neuronales artificiales.

### 3. Introducción

La farmacología basada en la fisiología (PBPK, por sus siglas en inglés, *Physiologically Based Pharmacokinetics*) ha emergido como un enfoque fundamental en el descubrimiento, desarrollo y uso de medicamentos. Los datos PBPK desempeñan un papel crucial al proporcionar una comprensión detallada de cómo los fármacos interactúan con el organismo, permitiendo una optimización más precisa de la dosis y una evaluación más completa de la seguridad y eficacia de los medicamentos. Aunque los datos PK (farmacocinética) tradicionales también son importantes, la incorporación de la fisiología en los modelos PBPK permite una predicción más precisa y personalizada de la farmacocinética de los fármacos en diferentes poblaciones y situaciones clínicas ([Jones and Rowland-Yeo, 2013](#)).

La importancia de los datos PBPK radica en su capacidad para simular y predecir cómo los fármacos se absorben, distribuyen, metabolizan y eliminan en el organismo. Estos datos proporcionan información detallada sobre los procesos fisiológicos y las interacciones farmacocinéticas que influyen en la exposición de un fármaco y su respuesta en el cuerpo. Al comprender estos aspectos, los investigadores y los profesionales de la salud pueden optimizar las dosis, minimizar los efectos adversos y mejorar la eficacia terapéutica.

#### 3.1 Principales Parámetros Farmacocinéticos (PK)

A continuación, se detallan los principales parámetros farmacocinéticos y su significado en el contexto del desarrollo y uso de fármacos. Para una mejor comprensión, se proporciona una tabla ([Tabla 1](#)) que resume estas métricas, incluyendo los términos en inglés y las unidades de medida comúnmente utilizadas.

**Tabla 1:** Parámetros Farmacocinéticos



Parámetros Farmacocinéticos	Término en Inglés	Unidad de Medida
C <sub>max</sub> (Concentración Máxima)	Maximum Concentration (C <sub>max</sub> )	mg/L or µg/mL
T <sub>max</sub> (Tiempo para alcanzar la Concentración Máxima)	Time to Maximum Concentration (T <sub>max</sub> )	horas
AUC (Área Bajo la Curva)	Area Under the Curve (AUC)	mg·h/L or µg·h/mL
Clearance (Eliminación)	Clearance (CL)	L/h or mL/min
V <sub>d</sub> (Volumen de Distribución)	Volume of Distribution (V <sub>d</sub> )	L or L/kg
Semivida (T <sub>1/2</sub> )	Half-life (T <sub>1/2</sub> )	horas
Biodisponibilidad	Bioavailability	%

- A. C<sub>max</sub>: Concentración máxima en plasma después de administración, indicando la exposición máxima al fármaco. Importante para evaluar eficacia y toxicidad.
- B. T<sub>max</sub>: Tiempo para alcanzar C<sub>max</sub>, indicando la velocidad de absorción y el inicio de acción terapéutica.
- C. AUC: Área bajo la curva del perfil de concentración-tiempo, medida de exposición total al fármaco en el tiempo, útil para comparar formulaciones y determinar dosis óptimas.
- D. Clearance: Tasa de eliminación del fármaco del cuerpo, crucial para determinar dosis e intervalos de administración, y para prevenir acumulación y toxicidad.
- E. V<sub>d</sub>: Volumen de distribución, medida de la distribución del fármaco en el organismo, influenciado por su lipofiliidad.
- F. Semivida (T<sub>1/2</sub>): Tiempo para que la concentración plasmática se reduzca a la mitad, útil para determinar la frecuencia de dosificación.
- G. Biodisponibilidad: Fracción del fármaco que llega a la circulación sistémica, comparada con una vía de referencia, indicando la disponibilidad del fármaco para la acción terapéutica.

Estos parámetros son fundamentales para el diseño de tratamientos, evaluación de fármacos y comprensión de sus interacciones con el organismo.

## 3.2 *Inteligencia Artificial*

La Inteligencia Artificial simula la inteligencia humana para realizar tareas complejas, utilizando algoritmos y modelos matemáticos para procesar datos, reconocer patrones, tomar decisiones y resolver problemas de manera autónoma.

En las últimas décadas, la IA ha transformado la ciencia y la tecnología, incluyendo la farmacología, mediante el análisis de grandes volúmenes de datos para generar conocimiento y facilitar decisiones informadas en la investigación y desarrollo de fármacos.

Dentro de la Inteligencia Artificial, el Aprendizaje Automático, utiliza algoritmos para que las máquinas aprendan de los datos y mejoren su rendimiento sin instrucciones específicas, adaptándose y mejorando con más datos ([Bote-Curiel et al., 2013](#)).

El *Machine Learning* incluye el aprendizaje supervisado, donde se entrena con datos etiquetados para predecir o clasificar nuevos datos, y el aprendizaje no supervisado, que busca patrones en datos no etiquetados, útil para descubrir relaciones en grandes conjuntos de datos, valioso en identificación de objetivos terapéuticos o agrupación de pacientes.

En farmacología, el *Machine Learning* ha acelerado el diseño de nuevos fármacos al analizar bases de datos de compuestos químicos para predecir su actividad biológica, reduciendo costos de ensayos experimentales ([Dara et al., 2021](#)).

El *Machine Learning* incluye el Aprendizaje Semi-Supervisado, que combina datos etiquetados y no etiquetados para mejorar el rendimiento del modelo, especialmente útil cuando los datos etiquetados son escasos o costosos ([A. Theissler and P. Ritzer, 2022](#)). Otros enfoques como el Aprendizaje por Refuerzo y el Aprendizaje por Transferencia no son relevantes en este contexto.

## 3.3 *Modelos de lenguaje*

Los modelos de lenguaje de IA, alimentados por grandes datos y potente procesamiento, han revolucionado la interacción con máquinas y entre sí, avanzando desde respuestas automáticas a conversaciones completas y contextuales. Este desarrollo, impulsado por la arquitectura de redes neuronales y



conjuntos de datos masivos, en los últimos años ha tenido un notable crecimiento ([Youssef, 2023](#)).

Los modelos de lenguaje han impactado múltiples industrias, incluyendo la biotecnología, donde han acelerado la identificación de patrones y relaciones en datos científicos y médicos, facilitando el desarrollo de nuevos medicamentos y terapias ([Weissler et al., 2021](#)).

En el campo de la informática, los modelos de lenguaje han mejorado la interacción humano-máquina, permitiendo a los asistentes virtuales comprender y responder comandos con precisión. Los sistemas de procesamiento de lenguaje natural han evolucionado para entender y responder preguntas complejas, mejorando la eficiencia de búsquedas y accesibilidad a la información. La evolución de estos modelos es el resultado de décadas de investigación en IA, heredando de modelos antiguos que inicialmente usaban reglas programadas manualmente, pero que eran rígidos y no manejaban la variabilidad del lenguaje humano.

La evolución de los modelos de lenguaje dio lugar a una nueva ola con la introducción de modelos estadísticos de lenguaje. Estos modelos, basados en la probabilidad y la estadística, marcaron avances significativos en la generación automática de texto y la traducción automática. A pesar de sus logros, estos modelos aún enfrentan limitaciones en la comprensión del contexto y la coherencia del lenguaje ([Naseem et al., 2021](#)).

### 3.4 Problemática

La falta de bases de datos completas y estandarizadas de datos PBPK representa un desafío significativo en la farmacología individualizada ([Wang et al., 2009](#)). Aunque existen diversas fuentes de datos PK y PBPK dispersas en la literatura científica ([Altman et al., 2012](#)), ([“PubChem”, 2008](#)), ([Papadatos et al., 2015](#)), ([Judson et al., 2008](#)) y en estudios preclínicos y clínicos, la recopilación y el acceso a estos datos pueden ser complicados y consumir mucho tiempo. Además, la falta de estandarización en la recopilación y presentación de los datos dificulta su integración y comparación entre diferentes estudios y fuentes de datos. ([Hernandez et al., 2021](#))

Las bases de datos completas y estandarizadas de datos PBPK son fundamentales en la farmacología individualizada, ofreciendo información valiosa para decisiones fundamentadas y personalizadas en medicamentos. Al sistemáticamente recopilar y estandarizar estos datos, se pueden identificar patrones y tendencias que explican cómo factores fisiológicos afectan la farmacocinética de los fármacos.

## 4. Objetivos

El objetivo principal de este trabajo es **desarrollar y aplicar una base de datos de parámetros farmacocinéticos** de distintos fármacos utilizando la farmacología basada en la fisiología (PBPK), **así como crear un modelo predictivo para dichos parámetros**. Este objetivo se logrará mediante la consecución de los siguientes objetivos específicos:

- En primer lugar, se busca **desarrollar una base de datos integral y estandarizada de datos farmacocinéticos** de distintos fármacos, tales como la Vida Media, Concentración Máxima, Tiempo hasta C<sub>max</sub>, Concentración Máxima en Estado Estable, Área Bajo la Curva, Aclaramiento, Biodisponibilidad y Dosis Máxima Tolerada. Estos datos se recopilaron e integraron desde diversas fuentes, como *PKPDAI*, *DrugBank* y otras fuentes relevantes. Este enfoque pretende abordar la problemática existente en la individualización de la farmacología, donde la falta de bases de datos completas representa un desafío.
- Además, se propone **implementar estrategias de normalización y estandarización de datos** con el objetivo de mejorar la coherencia en la presentación de datos, facilitando así su integración y comparación entre diferentes estudios y fuentes de datos dispersas en la literatura científica.
- Paralelamente, se pretende **optimizar las técnicas de minería de texto y modelos de lenguaje para lograr una extracción automatizada más precisa de información clave** como las especies involucradas en el estudio, enfermedades tratadas y valores farmacocinéticos asociados al estudio, contribuyendo así a una base de datos más rica y detallada.

- En consecuencia, se propone **el desarrollo de un modelo predictivo específico para la farmacocinética humana**, centrándose en la predicción de parámetros como la Vida Media a partir de los datos obtenidos en la base de datos creada. La finalidad es avanzar en la comprensión de la relación entre variables farmacocinéticas y proporcionar un modelo robusto para la predicción de estos parámetros utilizando descriptores moleculares. Cabe destacar que la misma metodología puede ser utilizada para predecir el resto de parámetros farmacocinéticos incluidos en la base de datos.
- Finalmente, se pretende **crear un repositorio *GitHub* que contenga toda la implementación del desarrollo de la base de datos y del modelo predictivo**, para que esté públicamente accesible y pueda ser compartido con la comunidad científica interesada en el desarrollo y la aplicación de la farmacología basada en la fisiología (PBPK) y en el modelado predictivo de parámetros farmacocinéticos, promoviendo así, la transparencia, el acceso abierto a los métodos y datos utilizados en la investigación y permitiendo que otros investigadores puedan revisar y utilizar el trabajo realizado para sus propios estudios y para el desarrollo de nuevas herramientas y metodologías en el campo de la farmacocinética.

En conjunto, la consecución de estos objetivos tiene como finalidad contribuir al avance de la farmacología individualizada, permitiendo una optimización más precisa de la dosis y una evaluación más completa de la seguridad y eficacia de los medicamentos.

## 5. Metodología

---

### 5.1 *Creación de la base de datos*

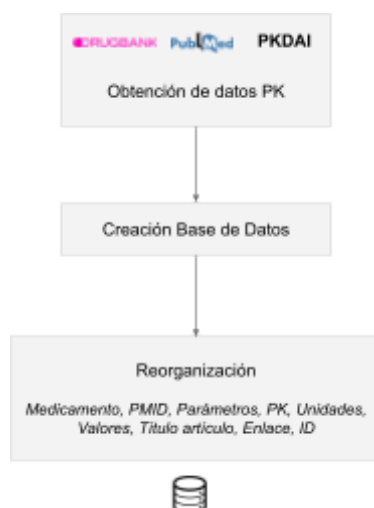
Con esta metodología se aborda la problemática de la falta de bases de datos completas y estandarizadas de datos PK, proponiendo una estrategia de extracción y análisis automatizado utilizando técnicas de minería de texto y modelos de

lenguaje avanzados. La estrategia de extracción y análisis se llevó a cabo utilizando el lenguaje de programación *Python* ([Guido, 1995](#)) y se utilizó una base de datos local *SQLite* (<https://docs.python.org/3/library/sqlite3.html>) para almacenar los datos.

### 5.1.1 Extracción de datos PK de PKPDAI y DrugBank

Se realizó un análisis exhaustivo de los datos de farmacocinética (PK) extrayendo información de las bases de datos *PKPDAI* (*Pharmacokinetics Parameter Data Aggregator and Identifier*) (<https://pkpdai.com/>) y *DrugBank* ([Wishart, 2006](#)). Se evaluaron diversas fuentes, considerando la idoneidad de programas y bases de datos disponibles, restricciones legales y de derechos de autor, así como costos y especificaciones técnicas para acceder y descargar datos (ver [Tabla 2](#) en Anexos). Se consideraron herramientas como *PK-DB* ([Grzegorzewski J et al., 2020](#)), *PK-Sim* y *MoBi* ([Eissing T et al., 2011](#)), *PLETHEM* ([Pendse et al., 2020](#)), *httk* ([Pearce RG et al., 2017](#)), *HESS* ([Sakuratani Y et al., 2013](#)), *DIDB* ([Hachad H et al., 2010](#)), *PK DDIs Database* ([Zhang S et al., 2022](#)), *DRUGBANK 3.0* ([Knox C et al., 2011](#)), *pkpdai* ([Hernandez F et al., 2021](#)), *DruMAP* ([Kawashima H et al., 2023](#)).

Se desarrolló una metodología eficiente en *Python* para extraer datos farmacocinéticos in vivo de la base de datos *PKPDAI*, enfocándose en medicamentos listados en *DrugBank*. Utilizando el paquete *requests*, se extrajeron y almacenaron los datos en una base de datos *SQLite* local. La organización de los datos se optimizó mediante una función que distribuye los atributos farmacocinéticos en columnas separadas, incluyendo identificadores como el PMID, detalles clave de farmacocinética, información del artículo, título, enlace y un ID único por entrada.

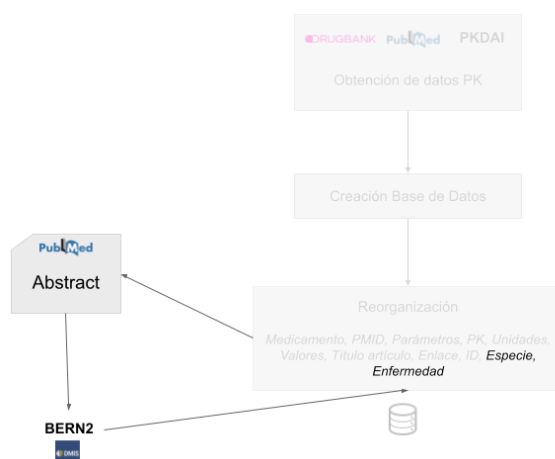


**Figura 1:** Esquema del Proceso de Obtención y Almacenamiento de Datos Farmacocinéticos

### 5.1.2 Minería de texto para extracción de "especie" y "enfermedad"

Al enfrentar el desafío de contextualizar adecuadamente los datos farmacológicos, se llevó a cabo un proceso de selección de modelos de lenguaje pre-entrenados para extraer la información de "especie" y "enfermedad" de los artículos científicos. Se evaluaron modelos como *flan-T5-Base* ([flan-t5-base](#)), *flan-t5-XXL* ([flan-t5-xxl](#)), *T5-base-for-BioQA* ([T5-base-for-BioQA](#)) y *flan-alpaca-large* ([flan-alpaca-large](#)), destacando el modelo *BERN2* ([Sung et al., 2022](#)) por su precisión en la tarea de *Reconocimiento de Entidades Nombradas (NER)* (ver [Tabla 3](#) de Anexos) .

La implementación de *BERN2* mejoró la precisión en la identificación de "especie" y "enfermedad" durante la extracción de información. Se creó un script en *Python* para automatizar la búsqueda en una base de datos *SQLite* local y realizar consultas a la API de *BERN2*, obteniendo datos en formato *JSON*. Estos datos se procesaron, normalizaron y limpiaron para eliminar duplicados y convertir el texto a minúsculas, asegurando la calidad y relevancia de la información al filtrar entradas sin probabilidad de acierto. Finalmente, se actualizaron los registros en la base de datos local, añadiendo columnas para "especie" y "enfermedad" asociadas a cada valor de PK. En la [Figura 2](#) se muestra una visión global llevada a cabo hasta ahora.



**Figura 2:** Creación de la base de datos de parámetros farmacocinéticos.

### 5.1.3 Extracción de Descriptores Moleculares

Una vez recopilados los datos farmacocinéticos relevantes, se procedió a la extracción de descriptores moleculares, fundamentales para comprender la naturaleza y las propiedades de las moléculas de los medicamentos en la base de datos. Los descriptores moleculares representan cuantitativamente diversas características moleculares, codificando información sobre la estructura y propiedades de una molécula en un formato numérico, siendo una herramienta crucial en quimioinformática. Para esta tarea, se optó por utilizar *Mordred*, una calculadora de descriptores moleculares que ofrece una solución robusta y eficiente (Moriwaki et al., 2018).

Mordred es capaz de calcular más de 1800 descriptores bidimensionales y tridimensionales.

**Tabla 4:** Ejemplos de Descriptores

Descriptor name	Number of descriptors (preset)
ABCIndex	2
AdjacencyMatrix	13
AtomCount	16

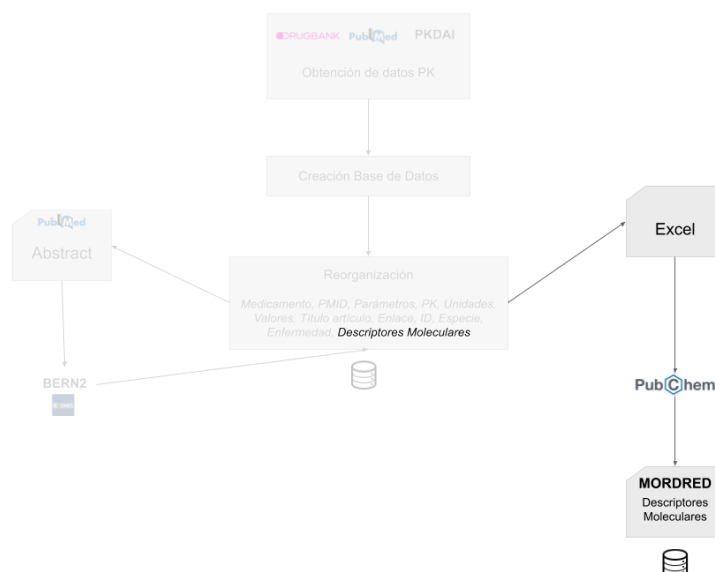
Autocorrelation	606
BondCount	9
Chi	56
DistanceMatrix	13
HydrogenBonda	2
MolecularId	12

Para obtener más detalles sobre estos descriptores, se puede consultar la tabla completa en la documentación del proyecto ([mordred-descriptor.github.io](https://mordred-descriptor.github.io)).

Para la extracción de descriptores moleculares con *Mordred*, se requiere una representación única de las moléculas en forma de *SMILES* (*Simplified Molecular Input Line Entry System*) ([Toropov et al., 2005](#)). Se implementó un procedimiento que utiliza un archivo Excel predefinido que contiene una lista de medicamentos asociados a sus respectivos *DRUGBANK\_ID*. A través del paquete *pubchempy*, se accedió a los *SMILES* correspondientes utilizando los identificadores "PubChem Compound" o "PubChem Substance". En casos donde solo se disponía de identificadores "PubChem Substance", se llevó a cabo un proceso adicional para obtener el CID estandarizado del compuesto, teniendo en cuenta que en algunos casos el CID estandarizado podría no existir.

Una vez obtenidos los *SMILES*, se procedió a la extracción de descriptores moleculares utilizando *Mordred*.

En la [Figura 3](#) se muestra una visión global llevada a cabo hasta ahora.



**Figura 3:** Creación de la base de datos de descriptores moleculares.

### 5.1.4 Limpieza de Datos

Se normalizaron y limpiaron los nombres de los parámetros farmacocinéticos en la base de datos de medicamentos para eliminar redundancias y asegurar coherencia. Este proceso incluyó identificar parámetros únicos por medicamento y analizar su frecuencia, abordando la variabilidad y ambigüedad en la nomenclatura para mejorar la calidad de los análisis.



**Figura 4:** Esquema del Proceso de Normalización y Limpieza

El proceso de normalización en la base de datos de medicamentos se llevó a cabo en *Python* para unificar los términos en la columna `Parameter`, mejorando la coherencia al estandarizar caracteres especiales, convertir el texto a minúsculas, ajustar plurales y eliminar guiones, espacios y caracteres no alfanuméricos, reduciendo 24,668 parámetros únicos a términos consistentes.



Tras la normalización, se obtuvo una lista reducida de parámetros únicos que facilitaron una representación homogénea en la base de datos, permitiendo un análisis detallado de la frecuencia y relevancia de cada parámetro.

Para simplificar la diversidad de términos en la base de datos de parámetros farmacocinéticos, se utilizaron diccionarios de sinónimos para agrupar términos comunes en categorías más amplias y coherentes. Estos diccionarios abarcaron los 300 primeros parámetros más recurrentes, facilitando una representación estructurada y uniforme de las características farmacocinéticas.

**Tabla 5:** Grupos de Sinónimos

Synonym Group	Synonym Terms
half-life	half-life, eliminationhalf-life, t12, terminalhalf-life, terminalhalf-life12, half-life, t12beta, t12alpha, half-life12, biologicalhalf-life, eliminationt12, eliminationhalf-life, eliminationhalf-life12, eliminationhalf-life, terminalhalf-life12, terminaleliminationhalf-life12, terminalplasmahalf-life, apparentterminalhalf-life, serumeliminationhalf-life, plasmahalf-life12, plasmahalf-life
c <sub>max</sub>	c <sub>max</sub> , maximumconcentration, peakplasmaconcentration, maximumplasmaconcentration, peakplasmalevel, peakconcentration, maximalplasmaconcentration, maximalconcentration, maximalplasmaconcentrationsc <sub>max</sub> , troughconcentration, plasma <sub>cmax</sub> , maximumplasmaconcentrationc <sub>max</sub> , maximalplasmaconcentrationc <sub>max</sub> , ct <sub>trough</sub> , troughplasmaconcentration
t <sub>1/2</sub> to c <sub>max</sub>	t <sub>1/2</sub> to maximumplasmaconcentration, t <sub>1/2</sub> to maximumconcentration, t <sub>1/2</sub> to c <sub>max</sub> , t <sub>1/2</sub> to c <sub>max</sub> t <sub>max</sub> , t <sub>1/2</sub> to peakplasmaconcentration, t <sub>1/2</sub> to peakconcentration, t <sub>1/2</sub> to each c <sub>max</sub> t <sub>max</sub> , peaktime
c <sub>ss</sub> max	steady-stateplasmaconcentration, c <sub>ss</sub> max, steady-stateplasmaconcentration
AUC	AUC, AUC0-infinity, AUC0-t, AUC0, AUCratio, AUC0-24, AUC0-24h, areaunderthecurve, areaundertheconcentrationtimecurve, areaundertheplasmaconcentrationtimecurve, areaundertheplasmaconcentrationtimecurveAUC, areaunderthecurveAUC, AUClast, AUCinf, AUC0-12h, AUCt, AUCinfinity, AUC0-8, AUC0-48
clearance	clearance, totalbodyclearance, plasma-clearance, systemicclearance, apparentclearance, apparentoralclearance, clearancerate, renalclearance, totalclearance, excretion, eliminationrateconstant, totalbodyclearancecl, renalclearancecl, clearancecl, metabolicclearance, bloodclearance, hepaticclearance, eliminationclearance, metabolicclearancerate, absorptionrateconstantka, nonrenalclearance, plasma-clearancecl
bioavailability	bioavailability, absolutebioavailability, oralbioavailability, relativebioavailability, systemicavailability, absoluteoralbioavailability, biologicalavailability, bioavailability, bioavailabilities, relativebioavailability, bioavailabilityf, absolutebioavailability, relativeoralbioavailability, bioavailable
maximal tolerated dose	MTD, maximal tolerated dose

La creación de diccionarios de sinónimos permitió agrupar términos relacionados en ocho categorías principales correspondientes a características específicas, logrando una representación más organizada y descriptiva. Esta normalización profunda sentó las bases para análisis avanzados en las etapas posteriores del estudio.

### 5.1.5 Normalización de Unidades y Valores

La normalización de unidades y valores es esencial para garantizar la coherencia y fiabilidad de los datos en preparación para análisis detallados. Se realizó una evaluación inicial de la variabilidad de unidades en cada grupo de sinónimos de parámetros, combinando herramientas de *SQL* y *Python* para comprender la diversidad de unidades y desarrollar estrategias efectivas de estandarización.

Para abordar la necesidad de comparabilidad entre medicamentos y evitar interpretaciones erróneas debido a unidades variables, se implementó una estrategia avanzada basada en un análisis detallado de las repeticiones de unidades. Se identificaron las unidades más representativas que abarcaban el 95% del total de repeticiones en cada grupo de sinónimos, estableciendo un conjunto de unidades clave para la normalización. Se creó un diccionario que asignaba un valor de conversión de 1 a cada unidad representativa, permitiendo la normalización coherente y estandarizada de las unidades menos frecuentes en relación con las más representativas.

Además, se llevó a cabo una normalización de los valores numéricos asociados a estas unidades, incluyendo limpieza inicial, conversión de palabras a números, y extracción de valores numéricos en diferentes formatos. Un diccionario de conversión se implementó mediante consultas *SQL* y manipulaciones en *Python* para asegurar una normalización homogénea en toda la base de datos, garantizando coherencia para análisis confiables.

En la segunda fase, se aplicó un diccionario de conversión a la base de datos mediante consultas *SQL* para reemplazar las unidades menos representativas en cada grupo por sus equivalentes en la unidad más representativa, utilizando los factores de conversión establecidos. Este proceso iterativo garantiza una normalización precisa y coherente en cada grupo de sinónimos.

Se llevó a cabo una validación exhaustiva para verificar la correcta aplicación de los cambios. Se compararon las unidades originales con las normalizadas mediante consultas de selección para evaluar la coherencia, asegurando que los valores asociados a las unidades también se normalizaran adecuadamente.

Para validar la efectividad de la metodología, se introdujeron dos nuevas columnas en la base de datos: `accepted_value` y `accepted_unit`. La columna `accepted_value` indicaba la validez de los valores después de la normalización, etiquetando como `True` aquellos con éxito y como `False` aquellos que no cumplieron con las reglas establecidas. De manera similar, la columna `accepted_unit` validaba las unidades asociadas a los valores, marcando como `True` las unidades normalizadas correctamente y como `False` aquellas que no pudieron convertirse efectivamente.

### 5.1.6 Integración de Parámetros Descriptivos

Se integraron parámetros descriptivos en la base de datos de medicamentos para presentar de manera más informativa las características farmacocinéticas. Se añadieron tres nuevas columnas: `mean` para la media, `std` para la desviación estándar y `median` para la mediana. Estas métricas se calcularon sólo para valores normalizados (`True` en las columnas `accepted_value` y `accepted_unit`).

La columna `mean` refleja la tendencia general de los datos, proporcionando una idea de la magnitud promedio de los parámetros. `std` almacena la desviación estándar, indicando la dispersión de los valores alrededor de la media. `median` representa el valor medio de los datos ordenados y no se ve afectado por valores extremos.

Estos parámetros descriptivos ofrecen un resumen cuantitativo y preciso de las propiedades farmacocinéticas de los medicamentos. Facilitan una evaluación completa de la distribución, dispersión y tendencia central de los parámetros, mejorando la utilidad de la base de datos para investigaciones futuras y análisis exploratorios.

## 5.2 Desarrollo del Modelo Predictivo

Después de recopilar y preparar datos de parámetros farmacocinéticos, se enfoca en desarrollar un modelo predictivo específico para la farmacocinética humana, centrándose en la vida media (*half-life*) y filtrando los datos para incluir sólo instancias relacionadas con la especie "humano".

El conjunto de datos se divide en entrenamiento y prueba para entrenar y evaluar un modelo de aprendizaje automático que utiliza descriptores moleculares como variables independientes para predecir la vida media. Esta elección se basa en la capacidad de los descriptores moleculares para representar características únicas de las moléculas y sus posibles relaciones con la farmacocinética.

La optimización del modelo implica técnicas como ajuste de hiperparámetros para mejorar la precisión y generalización del modelo, explorando diversas configuraciones.

### 5.2.1 Limpieza de datos

La limpieza de datos es un paso esencial en la preparación de conjuntos de datos para análisis y modelado, implicando la corrección o eliminación de datos incorrectos, corruptos, duplicados o incompletos. Este proceso incluye la eliminación de observaciones duplicadas o irrelevantes, corrección de errores estructurales y filtrado de valores atípicos. La limpieza de datos es crucial para asegurar la precisión de los modelos de *machine learning* y la toma de decisiones informadas.

La falta de limpieza de datos puede resultar en análisis inexactos y decisiones erróneas a largo plazo. Por lo tanto, es fundamental realizar una limpieza rigurosa y sistemática para garantizar la integridad y confiabilidad de los datos utilizados en el análisis y modelado ([Xu et al., 2016](#)).

Se emplearon dos enfoques diferentes: uno utilizando la media ("mean") como valor objetivo y otro replicando el procedimiento con el valor mediano ("median").

En la fase inicial de limpieza de datos, se eliminaron instancias con valores "nan", variables no numéricas y valores booleanos. Para mitigar la influencia de posibles

valores atípicos, se excluyeron aquellos correspondientes al primer cuartil de las variables objetivo.

En la construcción de conjuntos de entrenamiento y prueba, se asignó el 80% de los datos al conjunto de entrenamiento y el 20% restante al conjunto de prueba. Las variables independientes (X) incluyeron todos los descriptores, excepto `mean`, `std` y `median`. La variable dependiente (y) fue seleccionada como la variable objetivo correspondiente (`mean` en el primer enfoque y `median` en el segundo).

### 5.2.2 Normalización y Estandarización

En la fase siguiente, se realizó un análisis detallado de la normalización y estandarización de datos para asegurar la coherencia y comparabilidad de las variables.

La normalización de datos es crucial en el preprocesamiento para facilitar la comparación y análisis al ajustar los datos a un rango común. La estandarización, también conocida como normalización, implica ajustar los datos para tener una media de 0 y una desviación estándar de 1, beneficioso para algoritmos que asumen distribución gaussiana. Por otro lado, la normalización de bases de datos se refiere a reglas y técnicas para minimizar redundancia y prevenir anomalías en bases de datos relacionales.

La estandarización es esencial para garantizar la calidad y la integridad de los datos en análisis y modelado. Al homogeneizar los datos a un formato común, se facilita su interpretación y análisis, contribuyendo a decisiones informadas y resultados precisos ([Misra, 2020](#)).

Durante el desarrollo del modelo predictivo para la farmacocinética humana, se exploraron cuatro enfoques distintos de normalización y estandarización de datos. En el primer enfoque, se estandarizaron tanto las variables de entrada como las variables objetivo para que tuvieran una media de cero y una desviación estándar de uno. En el segundo enfoque, solo se estandarizaron las variables de entrada, manteniendo las variables objetivo en su escala original para facilitar la interpretación en esa escala. En el tercer enfoque, se aplicó la normalización *Min-Max* a las variables de entrada, dejando las variables objetivo en su escala

original para restringir las variables de entrada a un rango específico. En el cuarto enfoque, las variables de entrada permanecieron en su escala original, mientras que las variables objetivo se normalizaron con el método *Min-Max*.

La selección de la estrategia de normalización y estandarización se basó en la comparación de resultados utilizando métricas de evaluación como el error cuadrático medio (*MSE*) y el coeficiente de determinación ( $R^2$ ), que ofrecieron una evaluación cuantitativa de la precisión y capacidad predictiva de cada configuración.

### 5.2.3 Modelos Utilizados

Se utilizaron varios modelos para el análisis de datos, incluyendo la Regresión Lineal Múltiple (*RLM*), el Regresor de Bosque Aleatorio (*Random Forest Regressor*), la Regresión de Vector de Soporte (*SVR*) y hasta cinco configuraciones diferentes de Redes Neuronales Artificiales (*ANN*).

La Regresión Lineal Múltiple es un modelo estadístico versátil que evalúa las relaciones entre una variable dependiente continua y múltiples variables independientes, proporcionando una línea base comprensible para la predicción de parámetros farmacocinéticos. El Regresor de Bosque Aleatorio combina métodos de aprendizaje en conjunto con árboles de decisión para predecir valores numéricos, siendo útil para conjuntos de datos grandes y con alta dimensionalidad.

La Regresión de Vector de Soporte es un tipo de Máquina de Vector de Soporte (*SVM*) utilizado para regresión, buscando predecir valores continuos con funciones que pueden ser lineales o no lineales, capturando patrones complejos en los datos. Las Redes Neuronales Artificiales se inspiran en el funcionamiento del cerebro humano y son útiles para tareas como reconocimiento de patrones y aprendizaje no supervisado, ofreciendo la capacidad de capturar relaciones complejas entre descriptores moleculares y parámetros farmacocinéticos.

Se investigaron diversas configuraciones de Redes Neuronales Artificiales, modificando la arquitectura de capas ocultas, funciones de activación y tasas de aprendizaje para evaluar su capacidad de modelar relaciones no lineales y capturar patrones complejos en los datos.

Cada modelo fue elegido por su capacidad para abordar distintos tipos de datos y desafíos analíticos. Al experimentar con diferentes modelos y configuraciones, se logró evaluar su desempeño y determinar cuál ofrecía los resultados más precisos y relevantes para esta investigación.

### 5.2.4 Implementación de Librerías de Optimización

Se ha llevado a cabo una selección de librerías de optimización que apoyan la mejora y evaluación de los modelos.

- A. *Keras Tuner* es una herramienta que soporta la técnica de *Auto-Keras*, facilitando la búsqueda automatizada de hiperparámetros para modelos de Keras. Las estrategias utilizadas incluyen Random Search, Bayesian Optimization y Hyperband, cada una con un enfoque único en la exploración y explotación del espacio de hiperparámetros con el fin de optimizar los modelos ([Shawki et al., 2021](#)).
- B. La librería *FLAML (Fast and Lightweight AutoML)* se ha usado por su habilidad para aprovechar la estructura del espacio de búsqueda y facilitar una selección eficiente de algoritmos y configuraciones ([Salehin et al., 2023](#)). En esta investigación, *FLAML* se ha empleado como una solución rápida y de bajo costo computacional que facilita la experimentación con un amplio rango de algoritmos de aprendizaje automático.
- C. *MLJAR* es un paquete de *AutoML* en *Python* que opera de manera supervisada (<https://mljar.com/>). Se ha utilizado *MLJAR* para llevar a cabo optimizaciones de los modelos predictivos mediante métodos como el *Random Search* y otros algoritmos evolutivos.

### 5.2.5 Selección de Características y Modelos

Para mejorar los resultados de los modelos farmacocinéticos, se implementaron estrategias enfocadas en la selección de características y la optimización adicional de los modelos. A continuación se detallan estas metodologías:

- A. Selección de Características: Se utilizaron diversos enfoques para identificar las variables más influyentes en la predicción de parámetros

farmacocinéticos, como el método basado en  $k$  *features* en el que se seleccionaron 201 características para el valor *mean* y 152 para el valor *median*. También se probó usando una fórmula ponderada, con la cual se exploraron configuraciones alternativas utilizando  $\frac{\text{mean} * 2 + \text{mean} * 3}{2}$  para la selección de características. Otra técnica usada estuvo basada en un enfoque estadístico por el cual se seleccionaron las  $k$  características fundamentales que destacaron significativamente en un 90% respecto al resto (16 para la media y 7 para la mediana). Finalmente se aplicó el método de eliminación hacia atrás que consistió en un enfoque iterativo que eliminó gradualmente las características menos significativas según el *p-valor* establecido en 0.05, contribuyendo a refinar y simplificar el conjunto de características.

- B. Reducción de Dimensionalidad: Se exploraron técnicas como *Análisis de Componentes Principales (PCA)* y *Regresión Principal Componente (PCR)*, aplicadas junto con los métodos de selección de características mencionados y sin restricciones, es decir, utilizando todas las características disponibles.
- C. Normalización: Se probaron diferentes métodos de normalización, como estandarización y escalamiento *min-max*, para preparar los datos antes de su entrada a los modelos.

Estas estrategias permitieron afinar y mejorar los modelos farmacocinéticos al identificar las características más relevantes, reducir la dimensionalidad de los datos y normalizarlos adecuadamente para su análisis.

Tras la implementación de las estrategias mencionadas, se seleccionaron los modelos más robustos y efectivos. Se prestó especial atención a la comparación de métricas de rendimiento, como el *Error Absoluto Medio (MAE)* y la *Raíz del Error Cuadrático Medio (RMSE)*. Dos modelos destacados de redes neuronales y el modelo proporcionado por la librería *MLJAR* fueron las opciones finales para la predicción de parámetros farmacocinéticos.



## 6. Resultados y Discusión

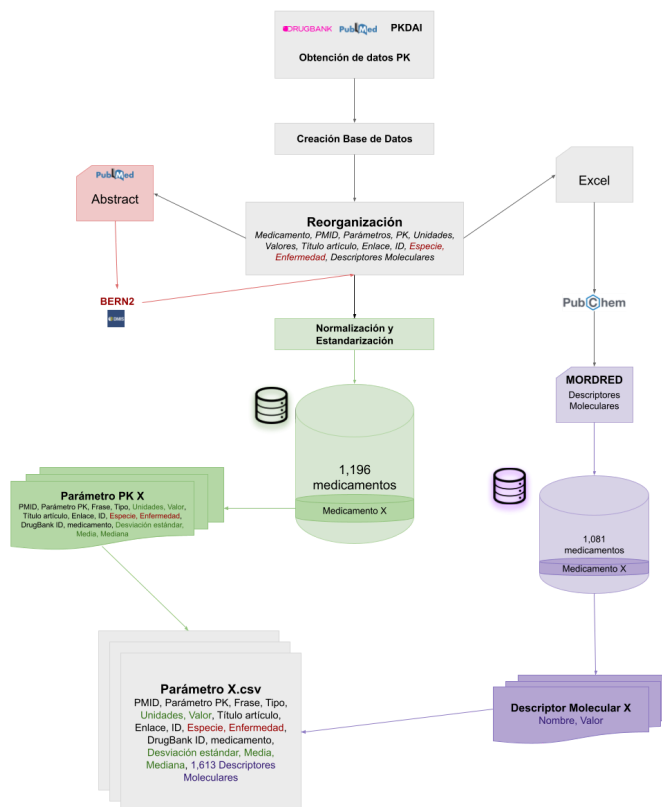
---

La implementación de la metodología propuesta en este estudio para la extracción de una base de datos farmacocinéticos y la creación de un modelo predictivo para estos parámetros (concretamente la vida media (*half-life*)) ha arrojado resultados satisfactorios y significativos. La primera etapa se centró en la extracción de información farmacocinética desde la base de datos *PKPDAI* hasta el 18 de julio de 2023, en relación con la base de datos *DrugBank*. En este sentido, se logró exitosamente la extracción de aproximadamente el 90% de los medicamentos presentes en *DrugBank* desde *PKPDAI*, con el restante 10% correspondiendo a fármacos ausentes en la base de datos *PKPDAI*. Esta fase fundamental ha permitido la construcción de una base de datos diversa en información farmacocinética, estableciendo así una base para análisis más profundos.

### 6.1 Creación de base de datos

A fecha actual, es decir, al 31 de julio de 2023, la base de datos contiene información detallada sobre 1,196 sustancias distintas.

A modo de esquema puede observar la [Figura 5](#), donde se resume todo el procedimiento llevado a cabo.



**Figura 5:** Esquema procedimental para la creación de la base de datos de descriptores moleculares y la base de datos de parámetros farmacocinéticos.

A continuación se detallan los resultados de las diferentes etapas.

### 6.1.1 Minería de Texto y Modelos de Lenguaje

En esta etapa del proceso, se emplearon técnicas de minería de texto y modelos de lenguaje avanzados para la extracción automatizada de información clave relacionada con "especie" y "enfermedad" asociadas a los valores farmacocinéticos. Entre los modelos de lenguaje evaluados, se estudiaron algunos especialmente destacados como los modelos *flan-T5-Base* y *flan-t5-XXL*, pertenecientes a la familia de modelos *T5*, eficaces en la comprensión de patrones complejos en el texto. También, el modelo *T5-base-for-BioQA*, especializado en la identificación de información relevante en el ámbito de la biología. Por otro lado, el modelo

*flan-alpaca-large*, basado en modelos *LLM*, también fue evaluado ya que es un modelo que proporciona de manera precisa información en textos científicos complejos.

### 6.1.2 Evaluación de Modelos y Precisión

Destacando en términos de precisión, el modelo *BERN2* se posicionó como una herramienta fundamental en este estudio. Basado en redes neuronales y enfocado en *Named Entity Recognition (NER)*, este modelo demostró una alta capacidad para reconocer entidades nombradas, como "especie" y "enfermedad". La evaluación de estos modelos se llevó a cabo comparando sus resultados con 25 papers previamente anotados manualmente, donde se especificaron las entidades de interés (ver [Tabla 6](#) en Anexos). Este enfoque permitió medir la precisión y el acierto en la identificación de estos elementos críticos para el análisis.

### 6.1.3 Cálculo de Precisión (Accuracy)

El cálculo de la precisión o *Accuracy* es una medida utilizada para evaluar la calidad de un modelo predictivo. Se basa en el porcentaje de predicciones correctas realizadas por el modelo sobre el total de predicciones realizadas. Para calcular la precisión de este modelo, se empleó la siguiente fórmula:

$$Accuracy = \frac{correct_{answers}}{incorrect_{answers} + correct_{answers}}$$

El número de preguntas correctamente respondidas, con las que se evalúa la precisión del modelo, se dividió entre la suma de las preguntas correctamente respondidas y las incorrectamente respondidas. Este cálculo se realizó tanto para determinar la enfermedad como para identificar la especie. Los resultados de ambas evaluaciones se sumaron y se dividieron entre 2, proporcionando así el valor de precisión del modelo *BERN2* y del resto de modelos (ver [Tabla 3](#) en Anexos).

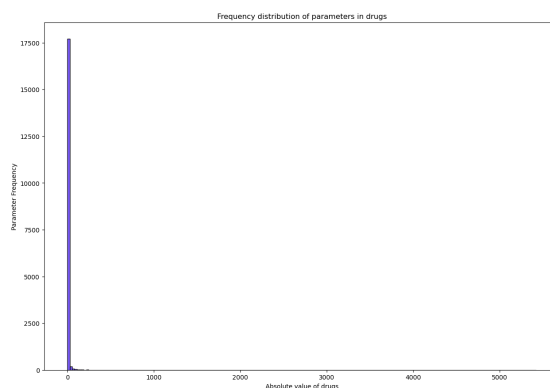
### 6.1.4 Descriptores Moleculares

Otro aspecto que mencionar fue la obtención de los *SMILES* asociados a los `drugbank_id`. Sin embargo, durante el proceso, se encontraron ciertas limitaciones.

En particular, se identificaron 11 medicamentos que carecían tanto de *CID* como de *SID*, y 104 medicamentos que no presentaban un *SMILE* asociado en la base de datos. A pesar de esto, se logró capturar exitosamente los *SMILES* de 1,081 medicamentos. Esta información resultó esencial para avanzar en el proceso de obtención de descriptores moleculares y comprender las características químicas y moleculares de las sustancias. Así pues, una vez que se obtuvieron los *SMILES*, el siguiente paso fue la extracción de descriptores moleculares. Para esto, se empleó el paquete *Mordred*, una herramienta para la generación de descriptores moleculares a partir de estructuras químicas. Los resultados de esta fase fueron notables: se logró extraer exitosamente descriptores moleculares para las 1,081 medicamentos. En total, cada una de estas medicamentos estuvo asociada con 1,613 descriptores moleculares únicos.

### 6.1.5 Normalización de Nombres de Parámetros

El proceso de normalización y limpieza de los nombres de los parámetros resultó de pasar de 24,668 parámetros únicos antes de la normalización (ver [Figura 6](#) en Anexos) a la obtención de 18,044 parámetros únicos después de la normalización. Entre las normalizaciones aplicadas se incluyó la conversión de caracteres especiales a su forma *ASCII* básica utilizando la función `unidecode`, la transformación de todos los caracteres a minúsculas, la eliminación de la letra "s" al final de las palabras, la supresión de guiones "-" y espacios en blanco, y finalmente, la eliminación de caracteres no alfanuméricos. En la [Figura 7](#), puede observar la distribución de parámetros únicos en los medicamentos. Obsérvese que la mayoría de parámetros únicos apenas lo tienen un 1% de los medicamentos.



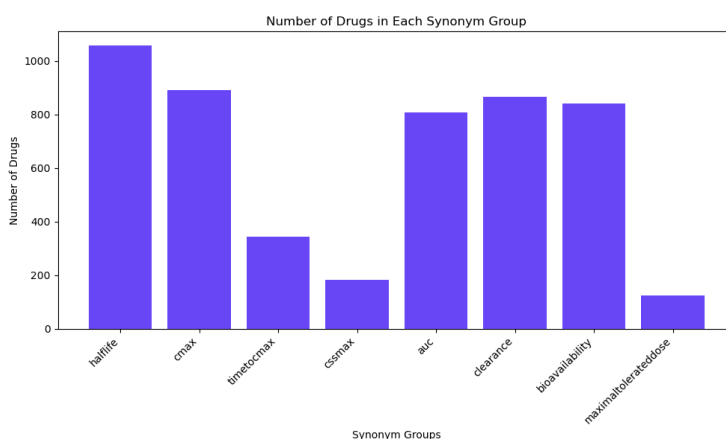
**Figura 7:** Distribución de frecuencias de parámetros en medicamentos

### 6.1.6 Grupos de Sinónimos

Esta cantidad, aún significativa de parámetros, proporcionó una visión integral de los aspectos farmacocinéticos que se estaban considerando y nos llevó a mejorar la estrategia de normalización de nombres de parámetros mediante la creación de diccionarios de sinónimos (ver [Tabla 5](#)). Tal y como muestra la [Figura 8](#) en los Anexos, el porcentaje de medicamentos que incluyeron al menos una columna con un grupo de sinónimos aumentó significativamente, alcanzando un 94.5%. Este resultado indicó que la mayoría de los medicamentos en la base de datos ahora estaban relacionados con al menos una característica farmacocinética coherente.

Por otro lado, sólo un 5.5% de las medicamentos no presentaron ninguna columna con un grupo de sinónimos correspondiente lo cual sugiere que la estrategia implementada resultó eficaz y resultó simplificar la representación de los datos.

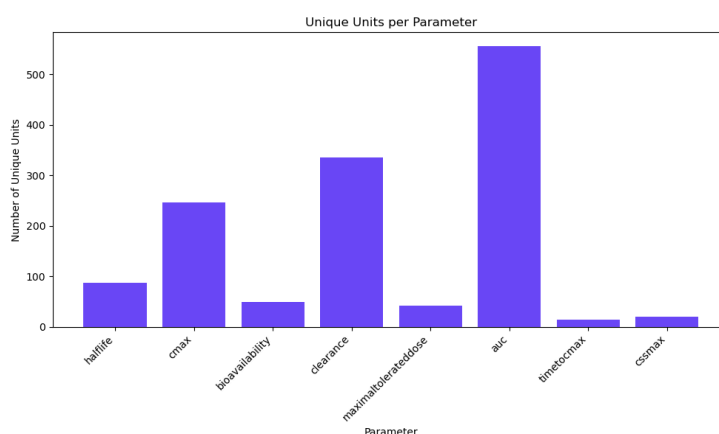
Tras incorporar los grupos de sinónimos se evaluó el número de medicamentos que contenían al menos un nombre en el grupo de sinónimos correspondiente, obteniéndose resultados satisfactorios. Por ejemplo, se observó que 1,057 medicamentos contenían al menos un nombre en el grupo de sinónimos de “half-life”. A continuación, en la [Figura 9](#) se muestran el número de medicamentos en cada uno de los grupos.



**Figura 9:** Número de medicamentos en cada grupo de sinónimos

### 6.1.7 Variabilidad de Unidades

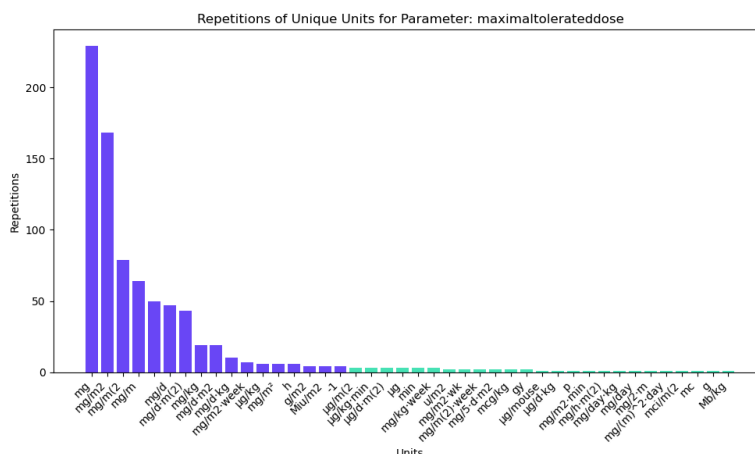
Como era de esperar, la evaluación de la variabilidad de unidades dentro de los grupos de sinónimos reveló una amplia diversidad para cada parámetro. Por ejemplo, en el grupo "half-life", se encontraron 87 unidades únicas, algunas de las unidades únicas identificadas incluyen:  $\mu M$ ,  $[g] / [h \cdot l]$ ,  $[l] / [h]$ ,  $ms$ ,  $1/[min]$ ,  $week$ , entre otras. Mientras que, por ejemplo, en el grupo "cmax" se identificaron 246 unidades únicas. Esta variabilidad se extendía a otros grupos como "bioavailability" con 50 unidades únicas y "clearance" con 336 unidades únicas. Estas cifras indican una diversidad significativa de unidades en cada grupo, lo que podría dificultar la comparación y el análisis directo de los datos. A continuación, en la [Figura 10](#), se muestran el número de unidades únicas por parámetro.



**Figura 10:** Número de unidades únicas por parámetro.

Con el fin de abordar este desafío, se diseñó un método que identificara las unidades que abarcaban el 95% del total de repeticiones en cada grupo de sinónimos. En esencia, se buscaban las unidades que eran más frecuentes y, por lo tanto, más representativas en el contexto de cada parámetro farmacocinético.

Por ello, se llevó a cabo un análisis automatizado y se realizó un reporte para cada uno de los 8 parámetros farmacocinéticos seleccionados para identificar las unidades pertenecientes a este 95%. Por ejemplo, a continuación, en la [Figura 11](#), se muestra la distribución de incidencias de unidades únicas para el parámetro "maxmaltolerateddose".



**Figura 11:** Repeticiones de unidades únicas para “maximaltolerateddose”

Para el resto de parámetros ver en el Anexo ([Figura 12](#), [Figura 13](#), [Figura 14](#), [Figura 15](#), [Figura 16](#), [Figura 17](#), [Figura 18](#)) .

Antes de aplicar los factores de conversión de unidades para estandarizar los valores en cada grupo de sinónimos, se llevó a cabo una depuración de los valores asociados a las unidades. Gracias a la implementación de operaciones de normalización en estos valores, se logró una drástica reducción en la cantidad de valores no numéricos que estaban inicialmente presentes en la base de datos. De un total inicial de 10,079 valores no numéricos, esta cifra se redujo significativamente a tan solo 277, lo que permitió que los datos estuvieran en un formato adecuado para su posterior conversión.

Luego de esto, se procedió a llevar a cabo la conversión manual de las unidades dentro de cada grupo de sinónimos que abarcara el 95% del total de repeticiones en el grupo. En este proceso, se designó la unidad con el mayor número de repeticiones en las instancias de cada grupo como la unidad de referencia (con un factor de conversión de 1). A continuación, se muestra parte de la [Tabla 7](#) de Anexos.

**Extracto Tabla 7:** Factores de conversión entre unidades en cada grupo de sinónimos (amarillo, unidad de referencia)

Parameter	Unit	Conversion Factor
half-life	h	1

half-life	min	0.0166667
c <sub>max</sub>	ng/ml	1
c <sub>max</sub>	µg/ml	1000
clearance	ml/kg·min	1
clearance	l/h·kg	16.66

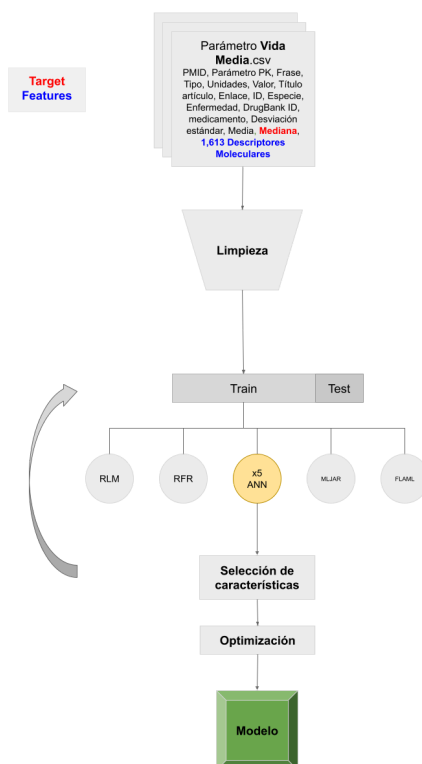
La tabla se presenta en el Anexo en la [Tabla 7](#).

Sin embargo, es relevante destacar que la conversión del 95% de las unidades no pudo ser alcanzada en todos los casos. Algunas conversiones, debido a su complejidad inherente, exigían información adicional no disponible en la base de datos. Factores como la falta de datos de la masa molecular, entre otros, obstaculizaron la conversión en algunos casos. Este conjunto de datos enriquecido y estandarizado proporciona una sólida plataforma para la realización de análisis farmacocinéticos avanzados y la identificación de patrones y tendencias significativas en el campo de la farmacología individualizada.

## 6.2 Desarrollo del Modelo Predictivo

El presente estudio se enfocó en el desarrollo de un modelo predictivo cuya finalidad fue predecir el tiempo de vida medio ( $T_{1/2}$ , *half-life*) en humanos a partir de los descriptores moleculares de los fármacos. A través de diversas etapas metodológicas, se buscó maximizar la precisión y generalización del modelo, abordando aspectos clave como la limpieza de datos, la normalización y estandarización, la selección de modelos y la optimización de hiperparámetros. A modo de esquema puede observar la [Figura 19](#), donde se resume todo el procedimiento llevado a cabo.





**Figura 19:** Esquema procedimental para la creación del modelo predictivo.

A continuación se detallan los resultados de las diferentes etapas.

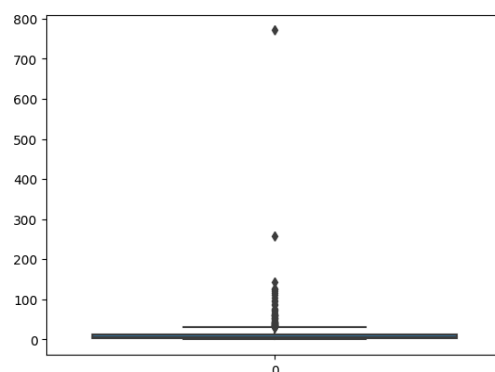
Se partió de un conjunto de datos que constaba de 723 fármacos. Este conjunto de datos se dividió en un conjunto de entrenamiento, que comprendía el 80% de los fármacos, y un conjunto de prueba, que comprendía el 20% restante. Esta división se realizó de forma aleatoria y estratificada para garantizar una representación equilibrada de los fármacos en ambos conjuntos. Así pues, el parámetro half-life o vida media se utilizó como variable objetivo (variable *target*) en el proceso de entrenamiento y validación del modelo.

### 6.2.1 Limpieza de Datos

En la fase inicial del proceso, se llevó a cabo la depuración de datos, eliminando instancias con valores "nan", variables no numéricas y valores booleanos.

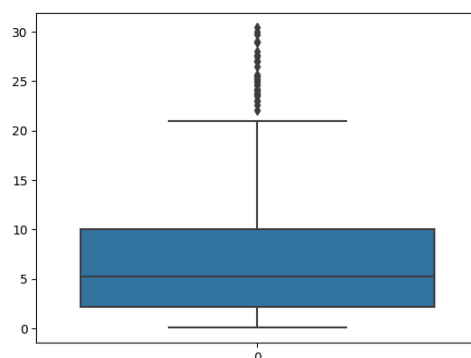
El número de filas antes de la eliminación fue de 723, y después de este proceso, se redujo a 709. Además, se identificaron 2 columnas de valores booleanos, las cuales

fueron eliminadas. Se procedió a generar un diagrama de caja utilizando los datos contenidos en la columna `median` del conjunto de datos, y se imprimió el valor de la mediana de dicha columna. La [Figura 20](#) revela valores distantes de la mediana.



**Figura 20:** Diagrama de caja de la variable target.

Estos análisis condujeron a la conclusión de que efectivamente se identificaron valores atípicos. Por consiguiente, se excluyeron observaciones correspondientes al primer cuartil de las variables objetivo para mitigar posibles efectos de valores atípicos ([Figura 21](#)).



**Figura 21:** Diagrama de caja de la variable target tras aplicar eliminación de Q1

Asimismo, se presentan dos gráficos comparativos (ver [Figura 22](#) y [Figura 23](#) en Anexo) en forma de histograma para la columna `median`, donde se superpusieron líneas verticales que representan la media y en un gráfico de histograma acumulado.

Posteriormente se aplicaron técnicas de preprocesamiento a través de aplicar diferentes combinaciones tanto de normalización como de estandarización a los datos. La normalización de los datos asegura que las características están en una

escala uniforme, evitando así que aquellas con magnitudes más grandes dominen la contribución al modelo. Se aplicó a un modelo estándar de regresión a las diferentes estrategias y se compararon métricas de evaluación, como el *Error Absoluto Medio* y la *Raíz del Error Cuadrático Medio*, de manera que se pudo constatar de manera iterativa que combinaciones arrojaban mejores resultados.

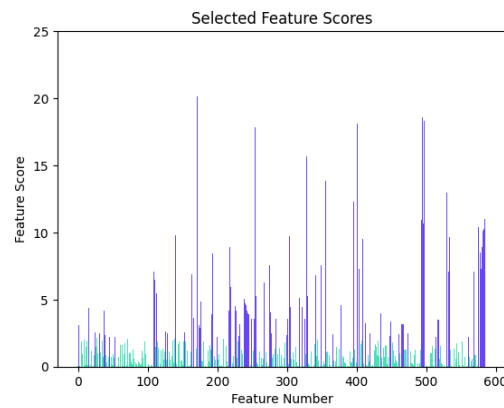
### 6.2.2 Modelos Utilizados

Se probaron diversos modelos con el propósito de explorar y evaluar el comportamiento del conjunto de datos. Entre estos modelos se encuentran la *Regresión Lineal Múltiple*, el *Regresor de Bosque Aleatorio*, la *Regresión de Vector de Soporte*, así como distintas configuraciones de *Redes Neuronales Artificiales* como ya se comentó.

Adicionalmente, se integraron herramientas de optimización proporcionadas por bibliotecas ya comentadas, tales como *Keras Tuner*, *FLAML* y *MLJAR*. Estas librerías no sólo incluyen la implementación de modelos, sino que también optimizan su rendimiento mediante estrategias como *Random Search* y *Bayesian Optimization*.

### 6.2.3 Selección de Características

Se aplicaron diversas estrategias para seleccionar características, incluyendo métodos basados en *k features*, los cuales se implementaron mediante el empleo de técnicas de correlación y análisis cuantitativo. En la primera etapa, se utilizó el método `f_regression` junto con el selector `SelectKBest` para calcular las puntuaciones de correlación entre las características de entrada y la variable de salida. Este enfoque inicial permitió identificar las características más influyentes en el contexto de un problema de regresión ([Figura 24](#)).



**Figura 24:** Gráfico de barras con puntuaciones de características: Destaca características clave en el análisis.

Posteriormente, se incorporó un análisis adicional para calcular la media de las puntuaciones de características y filtrar aquellas que superan este valor. Esta estrategia proporciona una visión más refinada al destacar las características que exhiben correlaciones significativamente por encima del promedio con respecto a la variable de salida. Estas estrategias proporcionan un análisis de las características, incorporando medidas de correlación, análisis cuantitativo y criterios de selección basados en la relevancia estadística.

Adicionalmente, se implementó la técnica de selección de características hacia atrás (*backward feature elimination*) utilizando *Recursive Feature Elimination (RFE)* en combinación con un modelo de regresión lineal. Este enfoque implica la eliminación iterativa de las características menos relevantes para mejorar la eficiencia del modelo seleccionando un número deseado de características (`k_features` obtenidas en el paso anterior). Esta técnica automatiza la selección de un conjunto específico de características relevantes para mejorar la capacidad predictiva del modelo de regresión lineal, identificando así las variables más influyentes en el proceso de predicción.

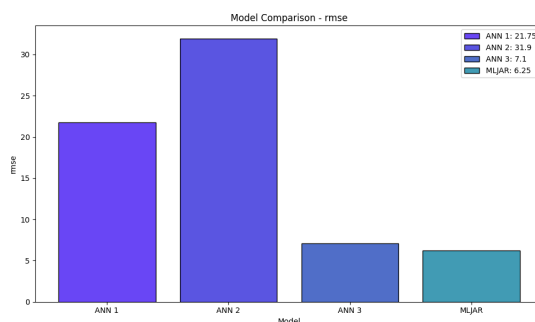
Finalmente, para probar otros enfoques, se probaron técnicas de reducción de dimensionalidad como el *Principal Component Analysis* y *Principal Component Regression*. *PCA* identificó las componentes principales que retienen la mayor variabilidad en los datos, mientras que *PCR* integró esta reducción de dimensionalidad en un modelo de regresión.

Estas métricas permiten una toma de decisiones informada sobre qué enfoques de selección de características y técnicas de reducción de dimensionalidad mejor se ajustaban a los objetivos.

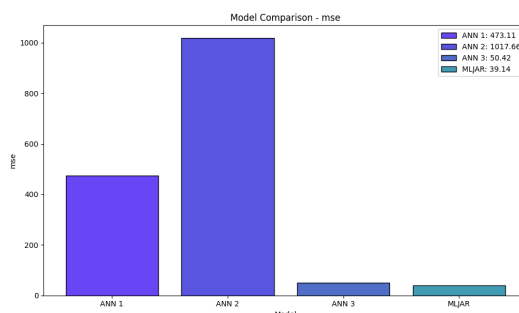
### 6.2.4 Modelos Finalistas

Tras la implementación de todas las estrategias, se destacan dos modelos (entre los cuatro modelos finalistas) como las opciones más robustas y efectivas para la predicción de parámetros farmacocinéticos. Sin embargo, es importante señalar que, a pesar de su destacado rendimiento, estos modelos no lograron proporcionar resultados completamente precisos y confiables. Este hallazgo sugiere la necesidad continua de mejoras en la recolección y tratamiento de datos en el campo de la farmacocinética humana.

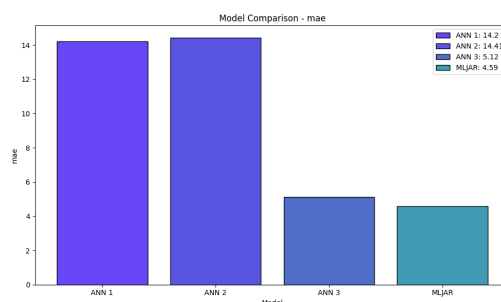
Se seleccionaron tres métricas clave: raíz del error cuadrático medio (*rmse*) (ver [Figura 25](#)), error cuadrático medio (*mse*) (ver [Figura 26](#)) y error absoluto medio (*mae*) (ver [Figura 27](#)). Para cada métrica, se crea un subgráfico que compara los resultados de los cuatro modelos finalistas. A continuación se muestra una comparativa entre los cuatro modelos finalistas:



**Figura 25:** Raíz del error cuadrático medio (*rmse*), modelos finalistas.

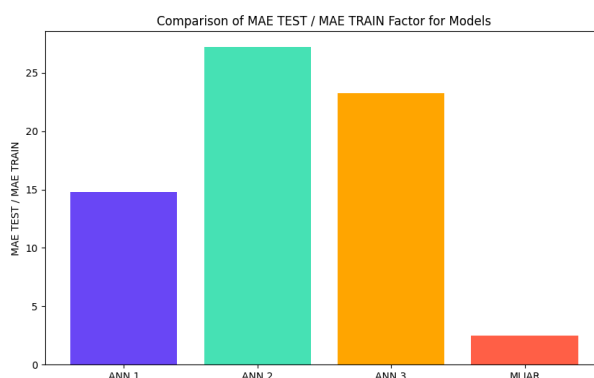


**Figura 26:** Error cuadrático medio (*mse*), modelos finalistas.



**Figura 27:** Error absoluto medio (mae), modelos finalistas.

Para comprender en detalle la optimización de los modelos durante la fase de entrenamiento y de prueba se comparó el *MAE* de la fase de prueba con la de entrenamiento (ver [Figura 28](#) en Anexo). Un análisis adicional permite ver de manera clara el factor resultante de dividir el *MAE* en el conjunto de prueba entre el *MAE* en el conjunto de entrenamiento para cada modelo. Este factor proporciona una perspectiva adicional sobre la coherencia y equilibrio en la capacidad predictiva de los modelos ([Figura 29](#)):

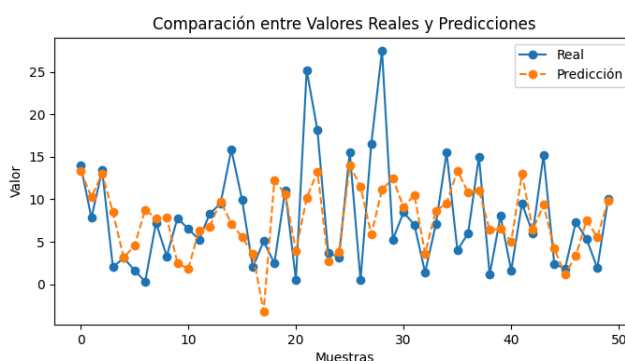


**Figura 29:** Factor Resultante MAE TEST / MAE TRAIN

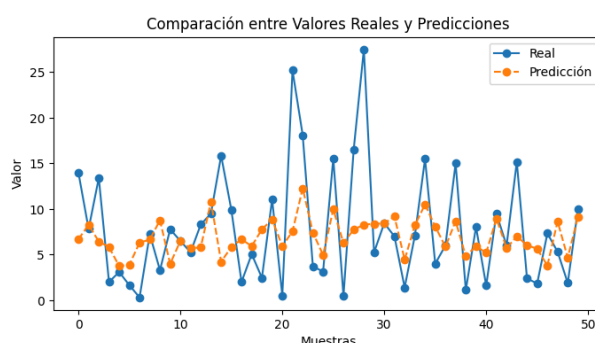
La elección de los dos últimos modelos (*ANN 3* y *MLJAR*) se basa en la evaluación de las métricas. En comparación con los dos primeros modelos (*ANN1* y *ANN 2*), los modelos *ANN 3* y *MLJAR* exhiben un rendimiento superior en términos de la métrica clave, el *Error Absoluto Medio*, tanto en los conjuntos de prueba como de entrenamiento. El *ANN 3* destaca con un *MAE* de 5.12 en el conjunto de prueba y 0.22 en el conjunto de entrenamiento, mostrando una tendencia a la sobreoptimización del modelo (ver [Figura 27](#)) pero demostrando una capacidad significativa para prever los parámetros farmacocinéticos. Además, su *RMSE* de

7.10 indica una buena capacidad de generalización y una menor dispersión de errores.

El *MLJAR*, aunque presenta un *MAE* ligeramente menor en el conjunto de prueba (4.59) que el *Modelo 3*, es importante destacar su *MAE* más alto en el conjunto de entrenamiento (1.83) lo que nos indica una menos sobreoptimización del modelo (ver [Figura 27](#)). Sin embargo, su *RMSE* de 6.26 sugiere que mantiene una capacidad de generalización robusta y ofrece una representación más equilibrada del rendimiento en ambos conjuntos. Para una comprensión detallada del rendimiento de los modelos seleccionados, *ANN 3* y *MLJAR*, en la predicción de parámetros farmacocinéticos, se presentan visualizaciones clave. La [Figura 30](#) y [Figura 31](#) muestran una comparativa entre los valores reales y las predicciones para las primeras 50 muestras del conjunto de prueba en el modelo *ANN 3* y *MLJAR* respectivamente. Cada punto en el gráfico representa un valor real (línea sólida) y su respectiva predicción (línea punteada).

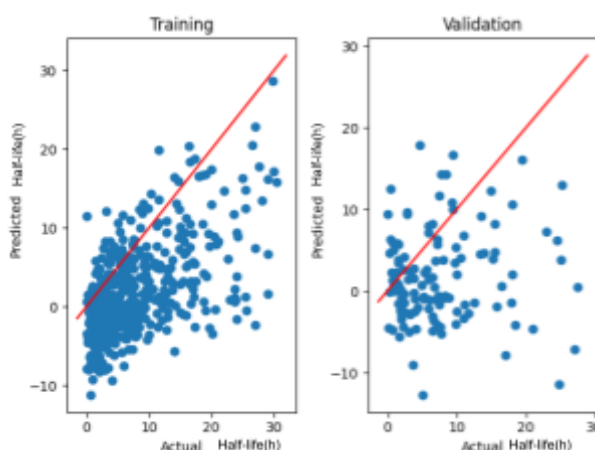


**Figura 30:** Comparación entre valores reales y predicciones - ANN 3



**Figura 31:** Comparación entre valores reales y predicciones - MLJAR

Finalmente, el modelo que visualmente mejor se ajusta es el modelo ANN 3. La [Figura 32](#) presenta scatterplots que contrastan los valores reales con las predicciones para los conjuntos de entrenamiento y validación para este modelo. Esta representación visual ofrece una evaluación inmediata de la precisión del modelo mediante la observación de la dispersión de puntos en relación con la línea de identidad.



**Figura 32:** Scatterplots de los valores reales con las predicciones - ANN 3

Considerando estos resultados, los modelos ANN 3 y MLJAR emergen como las opciones más sólidas, ofreciendo un equilibrio entre precisión en la predicción y capacidad de generalización.

## 7. Conclusiones

Se ha creado un [repositorio en GitHub](#) para albergar todo el trabajo realizado en este estudio. El repositorio contiene las bases de datos y el *pipeline* con los *scripts* necesarios para la creación de las bases de datos y del modelo predictivo, permitiendo así la extensión de su aplicación a otros parámetros farmacocinéticos recopilados en la base de datos u otras especies.

Puede acceder al repositorio en el siguiente enlace:

<https://github.com/pablosierrafernandez/FarmacoDB.git>



Esta iniciativa proporciona una plataforma para compartir el conocimiento y los avances logrados en este estudio, fomentando la colaboración y permitiendo que otros investigadores reproduzcan los resultados y contribuyan con nuevas mejoras en el campo de la farmacocinética.

## 7.1 Creación de la base de datos

La metodología propuesta en este estudio ha sido exitosa en la automatización de la extracción y análisis de datos farmacocinéticos. La base de datos actualmente cuenta con información detallada de 1,130 fármacos, representados por 100,359 registros de parámetros farmacocinéticos, permitiendo un análisis más detallado y preciso. La aplicación de técnicas avanzadas de minería de texto y modelos de lenguaje, como *flan-T5-Base* y *flan-t5-XXL*, ha sido crucial para la extracción automatizada de información clave sobre "especie" y "enfermedad" asociadas a los valores farmacocinéticos. La evaluación de modelos, especialmente el modelo *BERN2*, ha destacado en términos de precisión, siendo fundamental para reconocer entidades nombradas.

La obtención de datos moleculares, la extracción de descriptores moleculares y la normalización de nombres de parámetros podrían proporcionar una comprensión más profunda de cómo los valores farmacocinéticos variarán entre diferentes especies y condiciones fisiopatológicas. A pesar de algunas limitaciones en la obtención de datos como la masa molecular, se logró enriquecer la base de datos con información esencial.

La estrategia de normalización con grupos de sinónimos ha simplificado la representación de datos, mejorando la coherencia en la base de datos y permitiendo una visión integral de los aspectos farmacocinéticos considerados. La conversión de unidades ha abordado la diversidad de unidades en cada grupo de sinónimos, facilitando la comparación y el análisis directo de los datos.

## 7.2 Creación del modelo predictivo

El estudio se centró en el desarrollo de un modelo predictivo específico para la farmacocinética humana, con énfasis en el parámetro de vida media (*half-life*). El objetivo fue utilizar el modelo para predecir la vida media de medicamentos en humanos a partir de descriptores moleculares. Este enfoque ha revelado avances significativos en la comprensión de la relación entre variables farmacocinéticas.

La limpieza de datos, que incluyó la eliminación de instancias con valores "nan" y la exclusión de variables no numéricas y booleanas, proporcionó un conjunto de datos más consistente y redujo el impacto de valores atípicos en la variable objetivo *median*. Se aplicaron técnicas de preprocesamiento, normalización y estandarización para garantizar la uniformidad de las características y prevenir la dominación de características con magnitudes más grandes.

Se probaron diversos modelos, desde *Regresión Lineal Múltiple* hasta *Redes Neuronales Artificiales*, utilizando herramientas de optimización para mejorar su rendimiento. La selección de características se llevó a cabo mediante métodos basados en *k features*, análisis cuantitativo y técnicas de reducción de dimensionalidad como *PCA* y *PCR*.

La evaluación de modelos finalistas, incluyendo *ANN 3* y *MLJAR*, reveló que ambos destacan en términos de métricas clave como el *Error Absoluto Medio*, mostrando un equilibrio entre precisión y capacidad de generalización. Aunque el *ANN 3* exhibió una ligera sobreoptimización, su capacidad para prever parámetros farmacocinéticos fue ligeramente aceptable, respaldada por un *RMSE* indicando una buena generalización. *MLJAR*, con un *MAE* ligeramente menor en el conjunto de prueba, mostró una menor sobreoptimización y una representación equilibrada del rendimiento en ambos conjuntos.

Las detalladas visualizaciones que contrastan los valores reales con las predicciones de *ANN 3* y *MLJAR* respaldaron la elección de estos modelos. La dispersión de puntos en los scatterplots señaló una precisión, aunque no completamente exacta, proporcionando una aproximación significativa.

Los modelos *ANN 3* y *MLJAR* emergen como opciones robustas para la predicción de parámetros farmacocinéticos, ofreciendo un pipeline sólido para futuras

investigaciones y mejoras continuas en la recolección y tratamiento de datos en el campo de la farmacocinética humana y otras especies.

Estos resultados contribuyen al avance de la farmacología individualizada y sientan las bases para análisis más profundos en el campo de la farmacocinética.

## 8. Bibliografía

1. Altman, Russ B., et al. "Principles of Pharmacogenetics and Pharmacogenomics." *Cambridge University Press eBooks*, 2012, <https://doi.org/10.1017/cbo9781139051194>.
2. Bote-Curiel, Luis, et al. "Deep Learning and Big Data in Healthcare: A Double Review for Critical Beginners." *Applied Sciences*, vol. 9, no. 11, June 2019, p. 2331. <https://doi.org/10.3390/app9112331>.
3. Dara, Suresh, et al. "Machine Learning in Drug Discovery: A Review." *Artificial Intelligence Review*, vol. 55, no. 3, Aug. 2021, pp. 1947–99. <https://doi.org/10.1007/s10462-021-10058-4>.
4. A. Theissler and P. Ritzer, "EduML: An explorative approach for students and lecturers in machine learning courses," 2022 IEEE Global Engineering Education Conference (EDUCON), Tunis, Tunisia, 2022, pp. 921-928, doi: 10.1109/EDUCON52537.2022.9766719.
5. Eißing, Thomas. "A Computational Systems Biology Software Platform for Multiscale Modeling and Simulation: Integrating Whole-Body Physiology, Disease Biology, and Molecular Reaction Networks." *Frontiers in Physiology*, vol. 2, Jan. 2011, <https://doi.org/10.3389/fphys.2011.00004>.
6. Grzegorzewski, Jan, et al. "PK-DB: Pharmacokinetics Database for Individualized and Stratified Computational Modeling." *Nucleic Acids Research*, vol. 49, no. D1, Nov. 2020, pp. D1358–64. <https://doi.org/10.1093/nar/gkaa990>.
7. Guido, Van Rossum. *Python Tutorial*. 1 Jan. 1995, ir.cwi.nl/pub/5007.
8. Hachad, Houda, et al. "A Useful Tool for Drug Interaction Evaluation: The University of Washington Metabolism and Transport Drug Interaction Database." *Human Genomics*, vol. 5, no. 1, Jan. 2010, p. 61. <https://doi.org/10.1186/1479-7364-5-1-61>.
9. Hernandez, Ferran Gonzalez, et al. "An Automated Approach to Identify Scientific Publications Reporting Pharmacokinetic Parameters." *Wellcome Open Research*, vol. 6, Apr. 2021, p. 88. <https://doi.org/10.12688/wellcomeopenres.16718.1>.
10. Jones, Hm, and K. Rowland-Yeo. "Basic Concepts in Physiologically Based Pharmacokinetic Modeling in Drug Discovery and Development." *CPT: Pharmacometrics & Systems Pharmacology*, vol. 2, no. 8, Aug. 2013, p. 63. <https://doi.org/10.1038/psp.2013.41>.
11. Judson, Richard S., et al. "ACToR — Aggregated Computational Toxicology Resource." *Toxicology and Applied Pharmacology*, vol. 233, no. 1, Nov. 2008, pp. 7–13. <https://doi.org/10.1016/j.taap.2007.12.037>.
12. Kawashima, Hitoshi, et al. "DruMAP: A Novel Drug Metabolism and Pharmacokinetics Analysis Platform." *Journal of Medicinal Chemistry*, vol. 66, no. 14, July 2023, pp. 9697–709. <https://doi.org/10.1021/acs.jmedchem.3c00481>.
13. Knox, Craig, et al. "DrugBank 3.0: A Comprehensive Resource for 'Omics' Research on Drugs." *Nucleic Acids Research*, vol. 39, no. Database, Nov. 2010, pp. D1035–41. <https://doi.org/10.1093/nar/gkq1126>.
14. Misra, Biswapriya B. "Data Normalization Strategies in Metabolomics: Current Challenges, Approaches, and Tools." *European Journal of Mass Spectrometry*, vol. 26, no. 3, Apr. 2020, pp. 165–74. <https://doi.org/10.1177/1469066720918446>.
15. Moriwaki, Hiroto, et al. "Mordred: A Molecular Descriptor Calculator." *Journal of Cheminformatics*, vol. 10, no. 1, Feb. 2018, <https://doi.org/10.1186/s13321-018-0258-y>.
16. Naseem, Usman, et al. "A Comprehensive Survey on Word Representation Models: From Classical to State-of-the-Art Word Representation Language Models." *ACM Transactions on*

- Asian and Low-Resource Language Information Processing*, vol. 20, no. 5, June 2021, pp. 1–35. <https://doi.org/10.1145/3434237>.
17. Papadatos, George, et al. "Activity, Assay and Target Data Curation and Quality in the ChEMBL Database." *Journal of Computer-Aided Molecular Design*, vol. 29, no. 9, July 2015, pp. 885–96. <https://doi.org/10.1007/s10822-015-9860-5>.
  18. Pearce, Robert G., et al. "Htk: R Package for High-Throughput Toxicokinetics." *Journal of Statistical Software*, vol. 79, no. 4, Jan. 2017, <https://doi.org/10.18637/jss.v079.i04>.
  19. Pendse, Salil N., et al. "Population Life-course Exposure to Health Effects Model (PLETHEM): An R Package for PBPK Modeling." *Computational Toxicology*, vol. 13, Feb. 2020, p. 100115. <https://doi.org/10.1016/j.comtox.2019.100115>.
  20. "PubChem." *Springer eBooks*, 2008, p. 1599. [https://doi.org/10.1007/978-1-4020-6754-9\\_13806](https://doi.org/10.1007/978-1-4020-6754-9_13806).
  21. Sakuratani, Yuki, et al. "Hazard Evaluation Support System (HESS) for Predicting Repeated Dose Toxicity Using Toxicological Categories." *SAR And QSAR in Environmental Research*, vol. 24, no. 5, May 2013, pp. 351–63. <https://doi.org/10.1080/1062936x.2013.773375>.
  22. Salehin, Imrus, et al. "AutoML: A Systematic Review on Automated Machine Learning With Neural Architecture Search." *Journal of Information and Intelligence*, Oct. 2023, <https://doi.org/10.1016/j.jiixd.2023.10.002>.
  23. N. Shawki, R. R. Nunez, I. Obeid and J. Picone, "On Automating Hyperparameter Optimization for Deep Learning Applications," 2021 IEEE Signal Processing in Medicine and Biology Symposium (SPMB), Philadelphia, PA, USA, 2021, pp. 1-7, doi: 10.1109/SPMB52430.2021.9672266.
  24. Sung, Mujeen, et al. "BERN2: An Advanced Neural Biomedical Named Entity Recognition and Normalization Tool." *Bioinformatics*, vol. 38, no. 20, Sept. 2022, pp. 4837–39. <https://doi.org/10.1093/bioinformatics/btac598>.
  25. Toropov, Andrey A., et al. *Simplified Molecular Input Line Entry System (SMILES) as an Alternative for Constructing Quantitative Structure-property Relationships (QSPR)*. 1 Aug. 2005, [nopr.niscpr.res.in/handle/123456789/18068](https://nopr.niscpr.res.in/handle/123456789/18068).
  26. Wang, Zhiping, et al. "Literature Mining on Pharmacokinetics Numerical Data: A Feasibility Study." *Journal of Biomedical Informatics*, vol. 42, no. 4, Aug. 2009, pp. 726–35. <https://doi.org/10.1016/j.jbi.2009.03.010>.
  27. Weissler, E. Hope, et al. "The Role of Machine Learning in Clinical Research: Transforming the Future of Evidence Generation." *Trials*, vol. 22, no. 1, Aug. 2021, <https://doi.org/10.1186/s13063-021-05489-x>.
  28. Wishart, David S. "DrugBank: A Comprehensive Resource for in Silico Drug Discovery and Exploration." *Nucleic Acids Research*, vol. 34, no. 90001, Jan. 2006, pp. D668–72. <https://doi.org/10.1093/nar/gkj067>.
  29. Xu Chu, Ihab F. Ilyas, Sanjay Krishnan, and Jiannan Wang. 2016. Data Cleaning: Overview and Emerging Challenges. In *Proceedings of the 2016 International Conference on Management of Data (SIGMOD '16)*. Association for Computing Machinery, New York, NY, USA, 2201–2206. <https://doi.org/10.1145/2882903.2912574>.
  30. Youssef, A. "Unleashing the AI Revolution: Exploring the Capabilities and Challenges of Large Language Models and Text-to-image AI Programs." *Ultrasound in Obstetrics & Gynecology*, vol. 62, no. 2, June 2023, pp. 308–12. <https://doi.org/10.1002/uog.26297>.
  31. Zhang, Shijun, et al. "OUP Accepted Manuscript." *Database*, vol. 2022, Jan. 2022, <https://doi.org/10.1093/database/baac031>.

## 9. Autoevaluación

Durante el desarrollo de este Trabajo de Fin de Grado, se han explorado diversas áreas de la ciencia de datos y la inteligencia artificial, lo que ha resultado en un

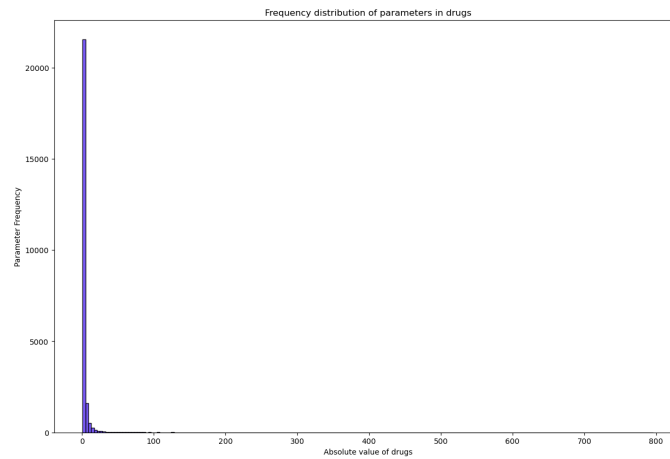
proceso de aprendizaje significativo y en el logro de varios objetivos personales y académicos. En cuanto a las expectativas iniciales, al comenzar este proyecto, esperaba adquirir un conocimiento más profundo sobre la aplicación de técnicas de aprendizaje automático en el campo de la farmacocinética. Además, aspiraba a mejorar mis habilidades en la limpieza y análisis de datos, así como en la implementación y evaluación de modelos predictivos. Durante el desarrollo del TFG, estas expectativas se cumplieron en gran medida.

Uno de los aprendizajes más destacados fue la comprensión del proceso completo de creación de un modelo predictivo, desde la recolección y limpieza de datos hasta la evaluación del rendimiento del modelo. A través de este proceso, pude aplicar los conocimientos teóricos adquiridos durante mi formación académica en un contexto práctico y real. Además, la realización de este trabajo me permitió desarrollar habilidades de investigación, resolución de problemas y toma de decisiones de forma independiente. A lo largo del proyecto, he fortalecido mi capacidad para abordar desafíos complejos y he mejorado mi destreza en el análisis crítico y la interpretación de resultados.

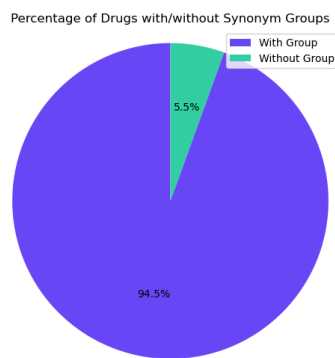
En resumen, el desarrollo de este Trabajo de Fin de Grado ha sido una experiencia enriquecedora que ha contribuido significativamente a mi crecimiento académico y profesional. He logrado alcanzar mis metas establecidas al inicio del proyecto y he adquirido habilidades valiosas que me serán útiles en mi futuro desarrollo profesional en el campo de la ciencia de datos y la inteligencia artificial.

## 10. Anexos

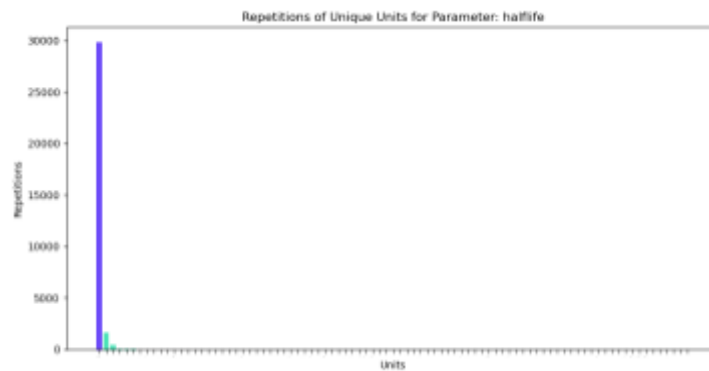
---



**Figura 6:** Frecuencia absoluta de distribución de parámetros únicos en las medicamentos



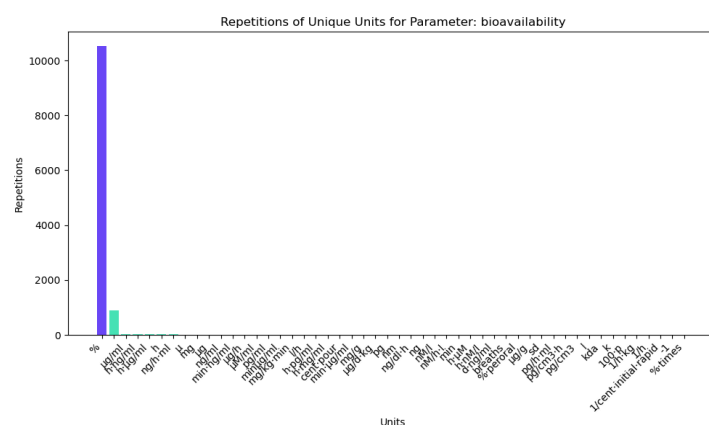
**Figura 8:** Porcentaje de medicamentos con y sin grupo de sinónimos



**Figura 12:** Repeticiones de unidades únicas para la vida media (half-life)



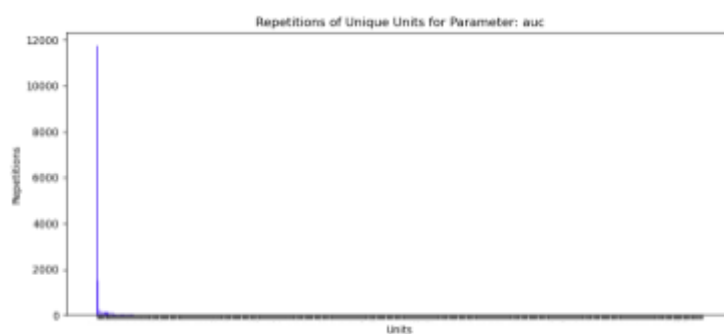
**Figura 13:** Repeticiones de unidades únicas para “cmax”



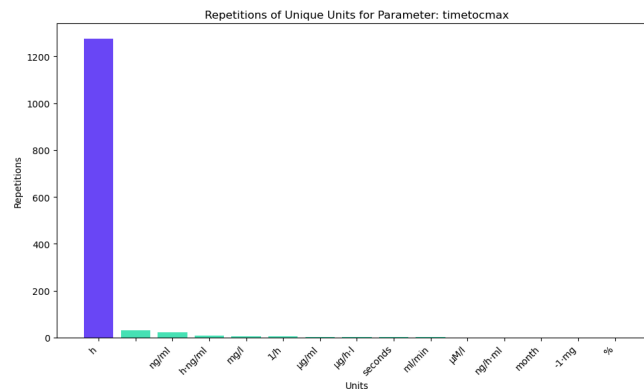
**Figura 14:** Repeticiones de unidades únicas para “bioavailability”



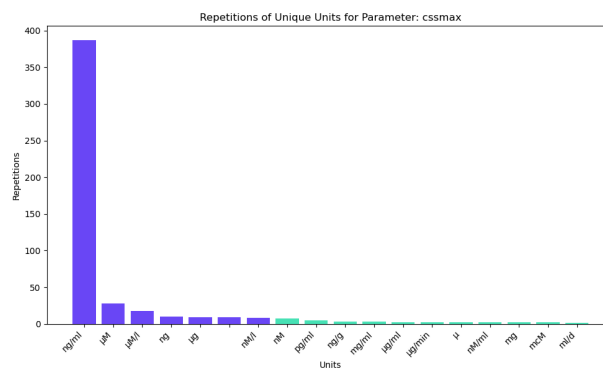
**Figura 15:** Repeticiones de unidades únicas para “clearance”



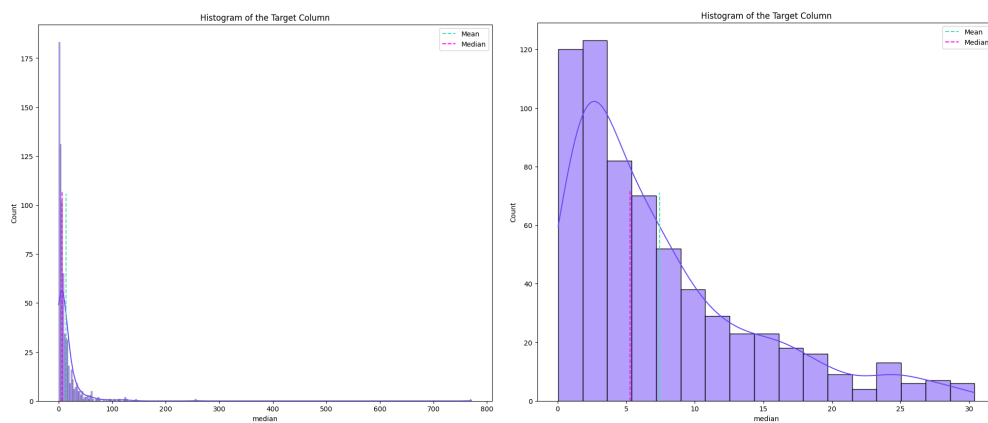
**Figura 16:** Repeticiones de unidades únicas para “auc”



**Figura 17:** Repeticiones de unidades únicas para “timetocmax”

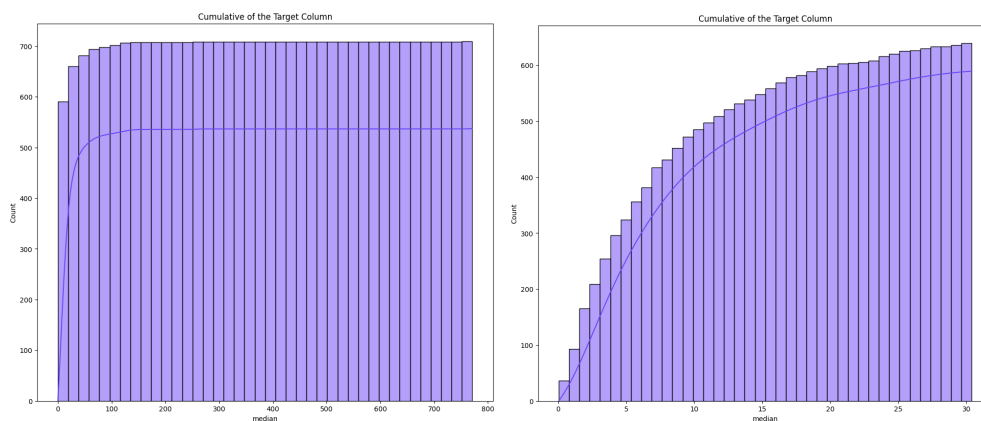


**Figura 18:** Repeticiones de unidades únicas para “cssmax”

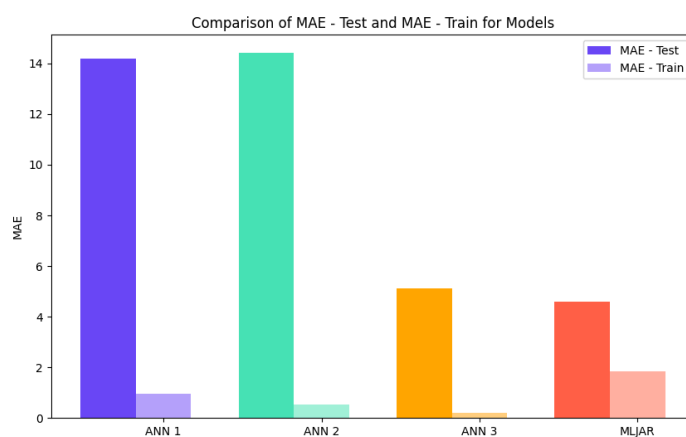


**Figura 22:** Histograma de la variable target con y sin eliminación de Q1.





**Figura 23:** Histograma acumulado de la variable target con y sin eliminación de Q1



**Figura 28:** Comparación de MAE - Test y MAE - Train

**Tabla 2:** Lista de Bases de datos PBPK

Database Name	Date of Publication	By Study?	Branch	Individuals	Raw PK Curve	Detailed Compartments	Comments	Type	License
Official name of DB	Date of first paper	If data is split by clinical trial	Detailed information of trials branches	Detailed information at the patient level	Raw datapoints of the PK curve	PK models are general, or by compartment			
PK-DB	2021	TRUE	TRUE	TRUE	TRUE	FALSE		Database (API)	GPLv3
PK-Sim and MoBi	2011	TRUE	TRUE	TRUE	TRUE	FALSE	It is not free, it is really software, although it has datasets of different species and individuals.	Software	Pay
PLETHEM	2019	FALSE	FALSE	FALSE	FALSE	FALSE	It is free, it is really software although I think it has a database.	Software	Free
httk	2017	FALSE	FALSE	FALSE	FALSE	FALSE		Database (downloadable) + R Library	GPL

HESS	2013	FALSE	FALSE	FALSE	FALSE	FALSE	It consists of 2 databases, one for toxicity and another for ADME information in rats and humans, but the individuality of the parameters is not specified anywhere.	Software + Database	The data can only be used in the software itself and not for other purposes
DIDB	2010	TRUE	TRUE	TRUE	FALSE	FALSE		Database	Pay
DDI-Corpus-Database	2022	TRUE	TRUE	TRUE	FALSE	FALSE		Database	Free
DrugBank 3.0	2011	FALSE	TRUE	FALSE	FALSE	TRUE		Database (API)	Pay
DruMap		FALSE	TRUE	FALSE	FALSE	TRUE		Database (API)	Pay
PKDocClassifier	2021	FALSE	FALSE	FALSE	FALSE	FALSE		PK paper classification model	Free

**Tabla 3:** Comparativa de Modelos

Model	Total # papers	# Correct	# not correct, also not wrong	# Wrong	Accuracy
<b>ozcangundes/T5-base-for-BioQA</b>					<b>0,5</b>
<i>What disease is the text about?</i>	6	2	0	4	0,33333333
<i>What specie of animal is involved?</i>	6	4	0	2	0,66666667
<b>google/flan-t5-base</b>					<b>0,789565217</b>
<i>What disease is the text about?</i>	25	21	0	4	0,84

What specie of animal is involved?	25	17	2	6	0,739130435
declare-lab/flan-alpaca-large					<b>0,793043478</b>
What disease is the text about?	25	19	2	4	0,826086957
What specie of animal is involved?	25	19	0	6	0,76
google/flan-t5-xxl					<b>0,833333333</b>
What disease is the text about?	25	18	1	6	0,75
What specie of animal is involved?	25	22	1	2	0,916666667
BERN2					<b>0,898333333</b>
What disease is the text about?	25	22	0	3	0,88
What specie of animal is involved?	25	22	1	2	0,916666667
timpal0l/mdeberta-v3-base-squad2					<b>0,708333333</b>
What disease is the text about or none?	25	18	1	6	0,75
What specie of animal is involved?	25	16	1	8	0,666666667

**Tabla 6:** Resultados anotados de 25 papers extrayendo la especie involucrada y la enfermedad.

Abstract ID (Pubmed)	What disease is the text about?	What specie of animal is involved?
1583404	deep vein thrombosis	patients(humans)
12560443	Hodgkin's lymphoma, cancer	patients/humans
24650013	female obesity	women
9496328	None	females

2525905	adenocarcinoma	BDF1 mice
30429607	Cancer	mice
12814453	None	healthy normotensive subjects
2186899	falciparum malaria	healthy adult male
33620754	tuberculosis	patients admitted for drug-resistant tuberculosis
24954342	pelvic floor muscle / Chronic pelvic pain / myofacial pain syndrome	female Wister rats
8872128	Uncontrolled activation of the tissue-factor (TF)-dependent	rabbit
12452736	myocardial infarction	patients
30884170	BW	patients
25212536	neuroblastoma	children
12123336	Neutropenia	Women
16151471	anemic / None	neonates
1908140	jugular vein thrombosis.	rabbit
26213478	anovulation	young female
10602568	myocardial infarction	patients
3612530	adenocarcinoma	rabbits and squirrel monkeys
10425368	none	rat
20042308	none	human
31093952	ischaemic stroke	human
23576486	locoregional or metastatic squamous cell head and neck cancer	human
12911582	pathogenesis of Type 1 von Willebrand disease (VWD),	humans

**Tabla 7:** Extracto de Tabla de Factores de conversión entre unidades en cada grupo de sinónimos (amarillo, unidad de referencia)

Parameter	Unit	Conversion Factor		Parameter	Unit	Conversion Factor
halflife	h	1		auc	mg/min·ml	16666.67
halflife	min	0.0166667		auc	mg·min/ml	16666.67
cmax	ng/ml	1		auc	min·ng/ml	0.01667
cmax	mg/ml	1000000		auc	mg·min/l	16.67
cmax	mcg/ml	1000		auc	d·ng/ml	24
cmax	ng/l	0.001		auc	ml·ng/h	1
clearance	ml/kg·min	1		auc	pg/h·ml	0.001
clearance	l/h·kg	16.66		auc	hng/ml	1
clearance	l/d·kg	0.694444444		auc	ng/d·ml	24
clearance	l/kg·mn	1000		auc	h·ug/ml	1000
bioavailability	%	1		auc	ng/h·l	0.001
maximaltolerateddose	mg	1		auc	ng/mlh	1
auc	h·ng/ml	1		auc	h·mcg/ml	1000
auc	ng/h·ml	1		auc	µg·hr/ml	1000
auc	µg/h·ml	1000		timetocmax	h	1
auc	h·µg/l	1		timetocmax	min	0.0167

