

Estrategias para el Tratamiento de “Missing data” en Ensayos Clínicos



Universitat
Oberta
de Catalunya

Álvaro Santos Gómez

MU Bioinf. y Bioest.
CLINICAL AND EPIDEMIOLOGICAL
STUDIES

Nombre Tutor/a de TF

Núria Pérez Álvarez

**Profesor/a responsable de la
asignatura**

Laia Subirats

18/06/2024





Esta obra está sujeta a una licencia de Reconocimiento-NoComercial [3.0 España de Creative Commons](https://creativecommons.org/licenses/by-nc/3.0/es/)

Copyright © 2024 Álvaro Santos Gómez.

Permission is granted to copy, distribute and/or modify this document under the terms of the GNU Free Documentation License, Version 1.3 or any later version published by the Free Software Foundation; with no Invariant Sections, no Front-Cover Texts, and no Back-Cover Texts.

A copy of the license is included in the section entitled "GNU Free Documentation License".

© (Álvaro Santos Gómez)

Reservados todos los derechos. Está prohibido la reproducción total o parcial de esta obra por cualquier medio o procedimiento, comprendidos la impresión, la reprografía, el microfilme, el tratamiento informático o cualquier otro sistema, así como la distribución de ejemplares mediante alquiler y préstamo, sin la autorización escrita del autor o de los límites que autorice la Ley de Propiedad Intelectual.

FICHA DEL TRABAJO FINAL

Título del trabajo:	<i>Estrategias para el Tratamiento de “Missing data” en Ensayos Clínicos.</i>
Nombre del autor:	<i>Álvaro Santos Gómez</i>
Nombre del director/a:	<i>Nuria Perez Alvarez</i>
Nombre del PRA:	<i>Laia Subirats</i>
Fecha de entrega (mm/aaaa):	<i>06/2024</i>
Titulación o programa:	MU Bioinf. y Bioest
Área del Trabajo Final:	<i>Estudios Clínicos y epidemiológicos</i>
Idioma del trabajo:	<i>Castellano</i>
Palabras clave	Missing data, Machine learning, Bioestadística

Resumen del Trabajo

This project focuses on evaluating various methods for handling missing data in the context of HIV clinical research. The dataset used originates from the Lake Study, which involves monitoring various clinical and laboratory parameters of patients living with HIV. The primary goal is to compare different imputation techniques to determine the most effective method for managing missing values, ensuring robust and reliable statistical analyses.

In this study, four main methods for handling missing data were implemented and evaluated:

- Row Deletion (Eliminación de Filas): Removing rows with missing values.
- Simple Imputation (Imputación Simple): Imputing missing values using the median.
- Modeling for Imputation (Modelado para Imputación): Predicting missing values using a preliminary model.
- Multiple Imputation (Imputación Múltiple): Using Iterative Imputer for multiple imputation.

These methods were evaluated using PCA and Random Forest model, and the performance was compared based on Root Mean Square Error (RMSE).

Abstract

This project focuses on evaluating various methods for handling missing data in the context of HIV clinical research. The dataset used originates from the Lake Study, which involves monitoring various clinical and laboratory parameters of patients living with HIV. The primary goal is to compare different imputation techniques to determine the most effective method for managing missing values, ensuring robust and reliable statistical analyses.

In this study, four main methods for handling missing data were implemented and evaluated:

- Row Deletion (Eliminación de Filas): Removing rows with missing values.
- Simple Imputation (Imputación Simple): Imputing missing values using the median.
- Modeling for Imputation (Modelado para Imputación): Predicting missing values using a preliminary model.
- Multiple Imputation (Imputación Múltiple): Using Iterative Imputer for multiple imputation.

These methods were evaluated using PCA and Random Forest model, and the performance was compared based on Root Mean Square Error (RMSE).

Índice

1.	Introducción.....	1
1.1.	Contexto y justificación del Trabajo	3
1.2.	Objetivos del Trabajo.....	4
1.3.	Impacto en sostenibilidad, ético-social y de diversidad	5
1.4.	Enfoque y método seguido	6
1.5.	Planificación del Trabajo.....	6
1.6.	Breve sumario de productos obtenidos	8
1.7.	Breve descripción de los otros capítulos de la memoria.....	8
2.	Estado del arte	10
3.	Material y métodos.....	12
4.	Resultados	16
5.	Conclusiones y trabajos futuros	32
6.	Glosario.....	38
7.	Bibliografía	39
8.	Anexos	42

1. Introducció

El manejo de los datos faltantes ('Missing data' en Inglés) es un desafío significativo en la investigación clínica, especialmente en estudios de enfermedades complejas (Fleming, 2011). Los datos faltantes, o "missings", se refieren a valores ausentes en un conjunto de datos debido a diversas razones, como errores en la recolección de datos, pérdida de muestras o problemas de seguimiento (Haukoos y Newgard, 2007). La presencia de datos faltantes puede llevar a sesgos en los resultados y afectar la validez de las conclusiones del estudio (Kenward, 2013).

Existen diferentes tipos de datos faltantes:

- 1- Completamente al Azar (MCAR): Los datos faltantes son completamente al azar si la probabilidad de que un dato esté ausente es independiente de cualquier otra variable. En este caso, los datos faltantes no introducen sesgo, pero pueden reducir la precisión del estudio (Fleming, 2011).
- 2- Al Azar (MAR): Los datos faltan al azar si la probabilidad de que un dato esté ausente depende de variables observables, pero no del valor de la variable faltante en sí misma. Este tipo de datos faltantes puede introducir sesgo si no se maneja adecuadamente, pero puede ser abordado mediante técnicas de imputación basadas en las variables observables (Fleming, 2011).
- 3- No al Azar (MNAR): Los datos no faltan al azar si la probabilidad de que un dato esté ausente depende del valor de la variable faltante. Este tipo de datos faltantes es el más complicado de manejar y puede introducir sesgos significativos (Fleming, 2011).

El tratamiento adecuado de los datos faltantes es crucial para asegurar la validez y la integridad de los análisis estadísticos. Diferentes técnicas, como la imputación y el análisis de casos completos, se utilizan para manejar este problema. La elección de la técnica adecuada depende del patrón y el tipo de datos faltantes.

Importancia del Tratamiento de Datos Faltantes

La ausencia de un manejo adecuado de los datos faltantes puede desvirtuar los resultados de un estudio, reduciendo la confiabilidad y la interpretabilidad de estos. Métodos simplistas como la imputación por la última observación disponible (LOCF), el análisis de casos completos o los análisis de sensibilidad extremos suelen ser inadecuados y pueden llevar a conclusiones incorrectas (Fleming, 2011). En estudios clínicos, la falta de datos puede ser especialmente problemática, ya que puede reducir la potencia del estudio y aumentar el riesgo de sesgo (Haukoos y Newgard, 2007).

Una forma efectiva de abordar los datos faltantes es prevenir su aparición mediante el diseño de protocolos de estudio robustos que maximicen la probabilidad de recolección de datos completos. Sin embargo, cuando los datos faltantes son inevitables, se deben emplear métodos de imputación racionales que consideren las suposiciones subyacentes y realicen análisis de sensibilidad para evaluar el impacto de diferentes métodos de imputación (Kenward, 2013).

Los métodos de imputación incluyen desde técnicas simples como la imputación media o regresión, hasta métodos más sofisticados como la imputación múltiple. La imputación múltiple es un enfoque estadístico robusto que permite manejar la incertidumbre asociada con los datos faltantes al generar múltiples conjuntos de datos imputados, analizar cada conjunto por separado y combinar los resultados para obtener inferencias válidas (Stack et al., 2018). Este enfoque es ampliamente recomendado en la investigación clínica debido a su capacidad para proporcionar estimaciones más precisas y reducir el sesgo.

Los datos faltantes en la investigación clínica pueden tener un impacto significativo en la validez de los resultados, afectando tanto la precisión como la interpretación de los hallazgos del estudio.

La pérdida de datos puede reducir el tamaño efectivo de la muestra, disminuyendo la potencia estadística del estudio y aumentando el error estándar de las estimaciones. Esto puede llevar a una menor capacidad para detectar efectos verdaderos, lo que incrementa la probabilidad de obtener resultados no concluyentes o falsos negativos (Fleming, 2011). Dependiendo del tipo de datos faltantes (MCAR, MAR, MNAR), los resultados pueden estar sesgados. Por ejemplo, si los datos faltan no al azar (MNAR), la ausencia de datos puede estar relacionada con el resultado que se está estudiando, lo que introduce un sesgo significativo en las estimaciones y puede llevar a conclusiones incorrectas (Kenward, 2013).

Además, los datos faltantes pueden llevar a la pérdida de información crucial sobre la variabilidad y las relaciones entre variables en el estudio, dificultando la comprensión completa del fenómeno que se está investigando y limitando la capacidad de generalizar los resultados a otras poblaciones o contextos (Haukoos y Newgard, 2007). La presencia de datos faltantes también puede afectar la validez externa de un estudio, es decir, la capacidad de generalizar los resultados más allá de la muestra estudiada. Si los datos faltan de manera sistemática en subgrupos específicos, los resultados pueden no ser representativos de la población completa, limitando la aplicabilidad de los hallazgos clínicos a otros grupos de pacientes (Stack et al., 2018).

1.1. Contexto y justificación del Trabajo

El Síndrome de Inmunodeficiencia Adquirida (Sida) es una enfermedad crónica, potencialmente mortal, causada por el virus de la inmunodeficiencia humana (VIH). Desde su identificación en la década de 1980, el Sida ha sido una de las principales preocupaciones de salud pública a nivel mundial. La infección por VIH afecta el sistema inmunológico, específicamente los linfocitos CD4, llevando a una disminución progresiva de la capacidad del cuerpo para combatir infecciones y otras enfermedades. Sin tratamiento, el VIH avanza hacia el Sida, una etapa avanzada de la infección en la que el sistema inmunológico está gravemente dañado.

El monitoreo de la infección por VIH y la progresión hacia el Sida se realiza mediante varios marcadores clínicos:

1. **Recuento de CD4:** Los linfocitos CD4 son células T auxiliares que desempeñan un papel crucial en el sistema inmunológico. El recuento de CD4 es un indicador esencial de la salud del sistema inmunológico en personas infectadas por el VIH. Un recuento bajo de CD4 indica una mayor progresión de la enfermedad y un mayor riesgo de infecciones oportunistas (Kagan et al., 2015).
2. **Carga Viral:** La cantidad de VIH en la sangre se mide mediante la carga viral, que es un marcador directo de la replicación del virus en el cuerpo. Una carga viral alta indica una mayor actividad viral y una mayor probabilidad de progresión de la enfermedad.

3. **Relación CD4/CD8:** Esta relación se utiliza para evaluar la recuperación inmunológica en pacientes con VIH. Un aumento en la relación CD4/CD8 se asocia con una mejor respuesta al tratamiento antirretroviral y una menor probabilidad de eventos clínicos graves (Bello et al., 2023).
4. **Otros Biomarcadores:** Además del recuento de CD4 y la carga viral, se monitorean otros biomarcadores como los niveles de colesterol (LDL, HDL) y la presencia de enfermedades asociadas para evaluar el estado general de salud y el riesgo cardiovascular en pacientes con VIH (Lembas et al., 2022).

1.2. Objetivos del Trabajo

Los objetivos del trabajo se centran en cuatro áreas principales: la revisión y comparación de métodos de imputación, el modelado predictivo, el impacto del tratamiento de los datos faltantes en la validez del estudio y la propuesta de mejoras y recomendaciones.

En primer lugar, se propone revisar la literatura existente sobre métodos de imputación de datos faltantes en estudios clínicos, comparar diferentes técnicas de imputación, como la imputación por la media, mediana, moda, imputación múltiple y métodos avanzados como la imputación iterativa, y evaluar la eficacia de cada método en términos de precisión y sesgo.

En el ámbito del modelado predictivo, el objetivo es entrenar modelos predictivos como Random Forest y HistGradientBoosting utilizando datos con diferentes métodos de imputación, y evaluar el desempeño de estos modelos en términos de precisión, sensibilidad, especificidad y otras métricas relevantes.

Además, se busca analizar cómo los diferentes métodos de imputación de datos faltantes afectan la validez general del estudio, discutiendo las implicaciones de los resultados en términos de sesgo y potencia estadística.

Finalmente, se pretende proporcionar recomendaciones sobre las mejores prácticas para manejar datos faltantes en estudios clínicos, y sugerir mejoras en la metodología de imputación basadas en los hallazgos del trabajo.

Estos objetivos están diseñados para abordar de manera integral los desafíos asociados con los datos faltantes en estudios clínicos, ofreciendo una evaluación detallada de las técnicas actuales y proponiendo soluciones prácticas para mejorar la validez y confiabilidad de los resultados de los estudios.

1.3. Impacto en sostenibilidad, ético-social y de diversidad

Sostenibilidad Medioambiental

El trabajo no impacta directa y severamente la sostenibilidad ambiental ni la huella ecológica. Siendo un proyecto de análisis de datos y modelado predictivo, la energía y los residuos generados no son un problema y solo se relacionan con el uso de recursos para realizar un proyecto en una computadora de sobremesa generalizada. Pero, de nuevo, también se debe equilibrar con la eficiencia de los algoritmos y modelos para no hacer la computación lenta y no consumir mucha energía. En cuanto a la ganancia, la mejora en las técnicas de imputación de valores faltantes también puede reducir la dependencia de repetir un estudio clínico, lo cual podría ser en línea con el principio de evitar y reducir el desperdicio.

Comportamiento Ético y Responsabilidad Social

La aplicación tiene un efecto ético-social en lo que respecta a la privacidad en la gestión de datos de salud. Al gestionar la información, la regulación actual con respecto a la protección de datos ha sido seguida para asegurar que la privacidad y la seguridad se logren al seguir las reglas y regulaciones establecidas. Aplicar técnicas más sofisticadas de imputación y predicción, en consecuencia, impulsará la práctica médica y contribuirá al ODS 3 (Salud y Bienestar) mediante el monitoreo y tratamiento mejorado de los pacientes con VIH. Asegurar una mayor transparencia y publicar tanto el código como la documentación en un repositorio de código abierto, lo que también permitirá una replicación equitativa del conocimiento para la comunidad científica y médica.

Diversidad, Género y Derechos Humanos

El proyecto es técnico y, por lo tanto, no tiene impacto directo en las preocupaciones de género, diversidad o derechos humanos. Pero, siendo un proyecto de datos relacionados con la salud, en cada uno de los análisis, se necesita poner a la diversidad de los pacientes en consideración de tal manera que los modelos sean inclusivos y estén creados de una manera que atienda a todas las divisiones de la población. La herramienta, si se aplica justamente, ayudaría a reducir las disparidades en el tratamiento de enfermedades y promovería el ODS 10: Reducción de las Desigualdades. El respeto de la diversidad y los derechos humanos en el desarrollo y la implementación de los modelos de imputación y predicción también es fundamental para asegurar que los beneficios del proyecto estén disponibles para toda la población objetivo, sin discriminación por género, raza, etnia y otros perfiles de la población.

El trabajo puede tener una gran influencia en el avance de la investigación clínica a través del uso de métodos innovadores de imputación de datos y modelos predictivos. El impacto positivo en la salud pública y la sociedad general está hecho máximo al asegurar la inclusión y la equidad, y al cumplir con todas las normas éticas y legales. Las líneas futuras de trabajo deben centrarse en los métodos para seguir mejorando y optimizando estas técnicas, con un enfoque continuo en la sostenibilidad, la responsabilidad social y la diversidad.

1.4. Enfoque y método seguido

Existen diversas estrategias para abordar este problema, y en este trabajo se ha decidido utilizar técnicas de tratamiento de datos faltantes ya existentes.

A continuación, se describen las posibles estrategias y la justificación para la elección de la estrategia seguida:

1. **Análisis de Casos Completos (Complete Case Analysis):** Esta técnica implica el análisis únicamente de los casos con datos completos. Aunque es simple de implementar, puede llevar a una reducción significativa del tamaño de la muestra y potencialmente introducir sesgos si los datos faltan no al azar (Haukoos y Newgard, 2007).
2. **Imputación Simple (Single Imputation):** Incluye métodos como la imputación de la media, imputación por regresión y la última observación (LOCF).
3. **Imputación Múltiple (Multiple Imputation):** Este enfoque genera varios conjuntos de datos imputados y combina los resultados de cada uno para obtener inferencias más robustas. (Fleming, 2011).
4. **Modelos Basados en Métodos de Aprendizaje Automático:** Utilizan algoritmos de aprendizaje automático para predecir y rellenar los valores faltantes. Estos métodos pueden ser muy eficaces para capturar relaciones complejas en los datos. Entre ellos, el uso de Random Forest ha demostrado ser una técnica robusta y precisa para la imputación de datos faltantes en estudios clínicos (Zhao et al., 2019).

1.5. Planificación del Trabajo

Para llevar a cabo este proyecto se necesitan varios recursos esenciales, incluyendo hardware adecuado como computadoras con suficiente capacidad

de procesamiento. También se requiere software especializado, incluyendo lenguajes de programación como Python y bibliotecas específicas como scikit-learn, pandas, numpy, matplotlib y seaborn para análisis de datos y visualización.

La planificación del proyecto se distribuye a lo largo de cuatro meses y se estructura en varias tareas clave, cada una con sus respectivos hitos.

La revisión bibliográfica se lleva a cabo en marzo y abril, enfocándose en buscar y revisar artículos relevantes y seleccionar metodologías adecuadas para la imputación de datos. En abril y mayo, se realiza el análisis y tratamiento de datos, incluyendo un Análisis Exploratorio de Datos (EDA), la identificación y manejo de valores faltantes, y la implementación de técnicas de imputación. En mayo, se lleva a cabo la comparación y visualización de resultados, donde se entrenan modelos predictivos y se evalúa su rendimiento, generando visualizaciones y análisis comparativos. En junio, se interpretan los resultados obtenidos, se redactan las conclusiones y recomendaciones, y se prepara la presentación final del trabajo, diseñando diapositivas y materiales de apoyo y ensayando la presentación. El diagrama de Gantt adjunto ilustra esta planificación temporal, marcando claramente los hitos parciales de cada tarea.

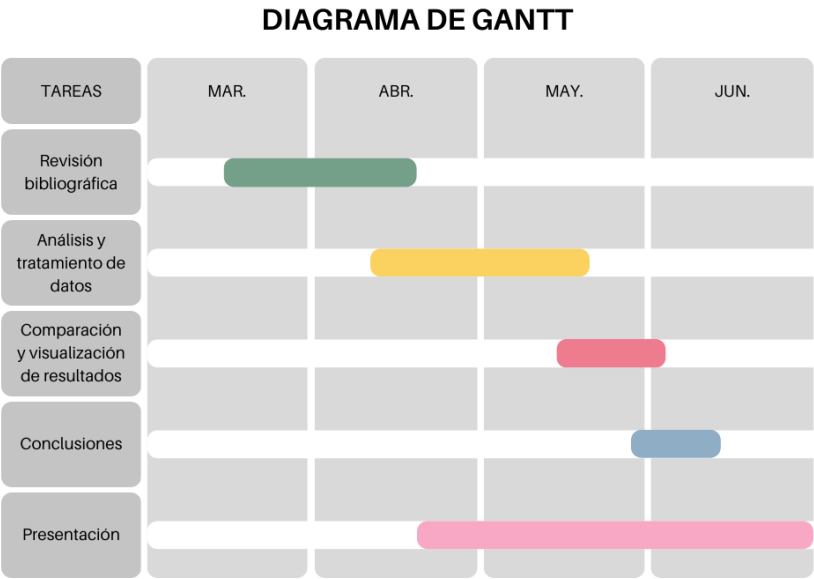


Fig 1. Diagrama de Gantt

A lo largo del proyecto, se identifican posibles riesgos que pueden afectar la planificación, como la falta de datos completos, problemas técnicos con el software, retrasos en la revisión bibliográfica, e inconsistencias en los datos. Para mitigar estos riesgos, se implementan diversas estrategias, como el uso de técnicas avanzadas de imputación y análisis de sensibilidad, mantener software actualizado y copias de seguridad. Esta planificación detallada

asegura que el proyecto se realice de manera organizada y eficiente, minimizando riesgos y garantizando la consecución de los objetivos planteados.

Durante el desarrollo del proyecto, surgieron varios imprevistos que obligaron a ajustar la planificación inicial. En particular, el análisis de datos y su tratamiento requirieron más tiempo del esperado debido a la complejidad de los datos y la necesidad de realizar un procesamiento exhaustivo para asegurar su calidad. Asimismo, la búsqueda bibliográfica se prolongó más de lo planeado, dado el volumen de literatura relevante que debía ser revisada para seleccionar las metodologías más adecuadas. Estos retrasos hicieron necesario reestructurar el cronograma, dedicando menos tiempo a la redacción de la memoria y acelerando las fases finales del proyecto. A pesar de estos desafíos, los ajustes realizados permitieron completar el proyecto con éxito, garantizando la calidad y robustez de los resultados obtenidos.

1.6. Breve resumen de productos obtenidos

A lo largo del proyecto se lograron varios productos clave. Se generó un conjunto de datos limpiado y preparado, imputando valores faltantes mediante imputación múltiple. También se entrenaron y evaluaron modelos predictivos, principalmente Random Forest y PCA, para mejorar la precisión de las predicciones. Además, se crearon visualizaciones detalladas, incluyendo gráficos y tablas, que facilitaron la comprensión de los resultados obtenidos.

El proceso metodológico se documentó exhaustivamente, abarcando desde la revisión bibliográfica hasta la implementación y evaluación de las técnicas de imputación. Además, he obtenido un documento Jupyter Notebook detallando el procesamiento y la limpieza de datos, así como todas las técnicas utilizadas en mi estudio. El código está completamente explicado y se incluyen los resultados obtenidos.

Para asegurar la transparencia y accesibilidad, se estableció un repositorio en GitHub con acceso libre, el cual incluye el código, los datos y la memoria del proyecto. El repositorio puede ser encontrado en la siguiente dirección: <https://github.com/Alvzant/Missings-data-in-clinical-trials.git>.

Los productos obtenidos se desarrollan más en el apartado 3.4.

1.7. Breve descripción de los otros capítulos de la memoria

El capítulo de Estado del Arte establece el contexto y justificación de nuestro enfoque, destacando la importancia de un tratamiento adecuado de los datos faltantes para asegurar la validez y confiabilidad de los resultados del estudio. Este capítulo revisa la literatura existente sobre métodos de imputación de

datos faltantes, y proporciona una base teórica sólida para las técnicas empleadas en nuestro trabajo.

El capítulo de Materiales y Métodos detalla el diseño del estudio y el conjunto de datos utilizado. Se describen los pasos seguidos para la limpieza y preparación de los datos, así como las estrategias empleadas para manejar los valores faltantes. Este capítulo es crucial para entender cómo se ha llevado a cabo el trabajo y proporciona una base para la reproducibilidad del estudio. Una limitación encontrada en este apartado fue la variabilidad y heterogeneidad de los datos originales, lo que complicó el proceso de limpieza y preparación, requiriendo un tiempo adicional para asegurar la consistencia de los datos.

En el capítulo de Resultados, se presentan los hallazgos del análisis de datos, incluyendo las evaluaciones de las diferentes técnicas de imputación. Se muestran gráficos y tablas que ilustran los resultados obtenidos, y se discute la efectividad de cada método en términos de precisión y sesgo. Este capítulo demuestra cómo las técnicas de imputación impactan en los resultados del modelo predictivo y proporciona una base para las conclusiones del estudio. Una de las limitaciones en esta sección fue el desafío de interpretar la gran cantidad de datos generados, lo cual demandó un análisis exhaustivo para extraer conclusiones significativas.

El capítulo de Conclusiones y Trabajos Futuros sintetiza los hallazgos del estudio, evaluando si se han cumplido los objetivos planteados inicialmente. Se discuten las implicaciones de los resultados y se sugieren áreas para futuras investigaciones. Entre las limitaciones encontradas se destaca el tiempo limitado para realizar un análisis más profundo de todas las técnicas de imputación disponibles, lo que sugiere la necesidad de estudios adicionales para explorar métodos más avanzados y sus aplicaciones en diferentes contextos clínicos.

2. Estado del arte

El tratamiento de datos faltantes es un desafío crucial en la investigación clínica. Los datos incompletos pueden afectar la validez de los resultados y conducir a conclusiones erróneas si no se manejan adecuadamente. En los últimos años, se han desarrollado y perfeccionado diversas técnicas para abordar este problema, desde métodos simples hasta enfoques más sofisticados.

1. Métodos Simples de Imputación

Los métodos simples de imputación son los más básicos y frecuentemente utilizados debido a su simplicidad y facilidad de implementación. Entre ellos se incluyen:

Imputación por la Media/Mediana:

Consiste en sustituir los valores faltantes con la media o la mediana de los datos disponibles. Aunque es fácil de implementar, este método puede introducir sesgos si los datos faltantes no son aleatorios (Little & Rubin, 2019).

Imputación por la Moda:

Similar a la imputación por la media, pero se utiliza la moda (el valor más frecuente) para los datos categóricos. Este método también puede ser problemático si los datos faltantes no son representativos del conjunto de datos (Graham, 2009).

2. Métodos Avanzados de Imputación

Los métodos avanzados buscan proporcionar imputaciones más precisas y menos sesgadas. Estos métodos incluyen:

Imputación Múltiple:

Este método implica crear múltiples conjuntos de datos imputados y combinarlos para obtener estimaciones más robustas. La imputación múltiple es considerada uno de los métodos más sólidos para manejar datos faltantes y es ampliamente utilizada en investigación clínica (Rubin, 1987).

Imputación Iterativa:

Utiliza modelos predictivos para estimar los valores faltantes en un proceso iterativo. Cada variable con datos faltantes se predice mediante un modelo que utiliza las demás variables como predictores. Este método puede capturar relaciones más complejas entre las variables (Van Buuren, 2018).

Métodos Basados en Machine Learning:

Algoritmos como K-Nearest Neighbors (KNN), Random Forest, y redes neuronales se utilizan para imputar datos faltantes basándose en patrones y relaciones en el conjunto de datos. Estos métodos pueden proporcionar imputaciones precisas, pero requieren más recursos computacionales y experiencia en su implementación (Troyanskaya et al., 2001).

Los datos faltantes pueden tener un impacto significativo en los estudios clínicos, afectando la validez interna y externa del estudio, así como la precisión de las estimaciones (Sterne et al., 2009). Un manejo inadecuado de los datos faltantes puede llevar a diversos problemas.

Si los datos faltantes no son aleatorios, las conclusiones del estudio pueden no ser representativas de la población objetivo, introduciendo sesgos en los resultados (Little & Rubin, 2019).

La reducción en el tamaño de la muestra efectiva debido a los datos faltantes puede disminuir la potencia estadística del estudio, aumentando la probabilidad de cometer errores tipo II (Graham, 2009).

Las conclusiones y recomendaciones basadas en datos incompletos pueden ser poco confiables y, en el peor de los casos, incorrectas, lo que puede tener implicaciones serias en el contexto clínico (Enders, 2010).

El presente trabajo se centra en la evaluación y comparación de diferentes métodos de imputación de datos faltantes en un conjunto de datos clínicos. La importancia de este estudio radica en varios aspectos clave que son fundamentales para mejorar la investigación en el ámbito clínico.

Primero, la mejora de la calidad de los análisis. Identificar y aplicar métodos de imputación más precisos permite mejorar la calidad de los análisis y las conclusiones derivadas de los estudios clínicos. Los resultados obtenidos con datos más completos y precisos ofrecen una base más sólida para la toma de decisiones médicas y el desarrollo de tratamientos.

Segundo, la reducción de sesgos. La implementación de técnicas avanzadas de imputación ayuda a minimizar los sesgos introducidos por los datos faltantes, lo cual mejora la representatividad y validez de los resultados. Esto es crucial para asegurar que las conclusiones del estudio sean aplicables a la población general y no solo a un subconjunto específico de pacientes.

Tercero, la aportación a la reproducibilidad. Documentar y estandarizar los métodos de imputación contribuye significativamente a la reproducibilidad de los estudios clínicos. Esto permite que otros investigadores puedan replicar y validar los resultados obtenidos, aumentando la credibilidad y el impacto del estudio.

Por último, el estudio ofrece propuestas para futuras investigaciones. Se sugieren diversas líneas de investigación y mejoras que podrían implementarse en futuros estudios para optimizar el tratamiento de datos faltantes y el modelado predictivo de CD4 en pacientes con VIH.

2.1. Diseño y Desarrollo del Trabajo

El diseño y desarrollo del trabajo implicó varias etapas clave:

Revisión de la Literatura: Revisión exhaustiva de métodos de imputación de datos faltantes en estudios clínicos.

Selección y Preparación del Conjunto de Datos: Exploración y limpieza de los datos obtenidos del Estudio Lake.

Implementación de Métodos de Imputación: Aplicación de diversas técnicas de imputación de datos faltantes.

Análisis y Evaluación: Aplicación de análisis estadísticos y modelado predictivo.

Documentación y Presentación de Resultados: Documentación detallada y presentación clara de los resultados obtenidos.

3. Material y métodos

3.2 Metodología

3.2.1. Revisión de la Literatura

Para establecer una base sólida, se revisaron artículos y libros clave sobre métodos de imputación de datos faltantes en estudios clínicos. Esta revisión

ayudó a identificar las técnicas más relevantes y las mejores prácticas para la imputación de datos.

3.2.2. Selección del Conjunto de Datos

El estudio utilizado en el análisis se basa en un ensayo clínico multicéntrico, abierto, prospectivo y aleatorizado, cuyo objetivo es evaluar la efectividad de dos regímenes de tratamiento antirretroviral en pacientes con VIH. Los tratamientos evaluados son:

Grupo 1: Abacavir 600 mg + Lamivudina 300 mg en pauta QD + Efavirenz 600 mg QD.

Grupo 2: Kaletra 400/100 mg BID.

El conjunto de datos del Estudio Lake [16] incluye información demográfica, clínica y de laboratorio de pacientes con VIH, recolectada en diversos hospitales de España. Este estudio me permitió analizar datos reales y comprender una situación compleja y previamente desconocida para mí: el VIH/SIDA y su tratamiento antirretroviral. La estructura de los datos presentaba una problemática compleja, especialmente en lo referente a los datos faltantes (missings). La limpieza inicial y la preparación del conjunto de datos se llevaron a cabo en varios pasos clave. Primero, se realizó una exploración inicial de las variables y su estructura, identificando la cantidad y ubicación de los datos faltantes. A continuación, se procedió a la limpieza de datos, eliminando columnas con más del 50% de datos faltantes para evitar sesgos significativos y asegurar la calidad de las imputaciones futuras.

La eliminación de variables con más del 50% de datos faltantes se fundamentó en varios factores importantes. La minimización del sesgo fue prioritaria, ya que las variables con tantos datos faltantes pueden introducir un sesgo significativo si se intentan imputar los valores faltantes, lo que podría llevar a conclusiones erróneas (Sterne et al., 2009).

Además, la calidad de la imputación es más efectiva cuando el porcentaje de datos faltantes es relativamente bajo, ya que con más del 50% de datos faltantes, las imputaciones pueden ser poco confiables y altamente especulativas (Graham, 2009). Mantener variables con un alto porcentaje de datos faltantes puede afectar la integridad y robustez del análisis, por lo que eliminarlas asegura que el análisis se base en datos más completos y representativos, mejorando así la validez interna del estudio. Por último, eliminar estas variables simplifica el análisis, reduciendo la complejidad y facilitando la interpretación de los resultados, lo que permite enfocarse en las variables más relevantes y completas.

Este enfoque asegura que el análisis se realiza con un conjunto de datos más robusto y confiable, mejorando la calidad y precisión de los resultados obtenidos, lo que es crucial en el contexto de estudios clínicos sobre VIH/SIDA. La experiencia de trabajar con datos reales y enfrentar los desafíos asociados a la complejidad de los datos faltantes ha enriquecido significativamente mi comprensión de este campo y ha subrayado la importancia de un manejo riguroso y sistemático de los datos en la investigación clínica.

3.2.3. Implementación de Métodos de Imputación

Se seleccionaron e implementaron varios métodos de imputación de datos faltantes utilizando Python y bibliotecas estadísticas como scikit-learn. Los métodos seleccionados se basaron en su relevancia y eficacia reportada en la literatura.

Imputación por la Media, Mediana y Moda

Estos métodos básicos sirvieron como referencia inicial. La fórmula para la imputación por la media es:

1. Imputación por la Media, Mediana y Moda: Uso de técnicas básicas como referencia(Little & Rubin, 2019).
2. Imputación Múltiple: Generación de múltiples conjuntos de datos imputados y combinación de resultados para obtener estimaciones robustas (Little & Rubin, 2019).
3. Imputación Iterativa: Utilización de modelos predictivos para estimar valores faltantes en un proceso iterativo (Van Buuren, 2018).
4. Métodos Basados en Machine Learning: Algoritmos como K-Nearest Neighbors (KNN) y Random Forest para imputar datos faltantes basándose en patrones y relaciones en el conjunto de datos (Troyanskaya et al., 2001; Zhao et al., 2023).

3.2.4. Análisis y Evaluación

Para evaluar el impacto de los métodos de imputación, se realizaron los siguientes análisis:

1. Análisis de Componentes Principales (PCA): Evaluación de la estructura de los datos antes y después de la imputación para entender el impacto en la reducción de dimensionalidad (Sterne et al., 2009).

2. Modelado Predictivo: Entrenamiento de modelos predictivos utilizando los datos imputados y evaluación de su desempeño en términos de precisión y otras métricas relevantes (Zhao et al., 2023).

3.2.5. Documentación y Presentación

Se documentaron todos los pasos y decisiones tomadas durante el desarrollo del trabajo. Se generaron gráficos y tablas descriptivas para presentar los resultados de manera clara y comprensible. Se utilizó Python y sus bibliotecas, tales como pandas, numpy, matplotlib, y seaborn para la visualización de los datos y resultados.

3.3. Alternativas Consideradas y Decisiones Tomadas

Durante el desarrollo del trabajo, se consideraron varias alternativas metodológicas y se tomaron decisiones clave. Una de estas decisiones fue la selección de métodos de imputación, optando por métodos simples y avanzados para proporcionar una evaluación comparativa exhaustiva. También se decidió eliminar columnas con más del 50% de datos faltantes durante la limpieza de datos, con el fin de evitar sesgos significativos (Sterne et al., 2009). Asimismo, se seleccionaron el análisis de componentes principales (PCA) y el modelado predictivo como las técnicas principales para evaluar el impacto de la imputación de datos (Zhao et al., 2023).

Los criterios utilizados para tomar estas decisiones incluyeron una revisión exhaustiva de la literatura existente, la naturaleza de los datos disponibles y la necesidad de proporcionar estimaciones precisas y robustas. La revisión de la literatura permitió identificar las mejores prácticas y metodologías más adecuadas, mientras que la evaluación de la naturaleza de los datos ayudó a determinar la viabilidad y pertinencia de las técnicas seleccionadas. Además, la necesidad de garantizar la precisión y robustez de las estimaciones fue fundamental para guiar la selección de las metodologías empleadas.

3.4. Descripción de los Productos Obtenidos

A lo largo del desarrollo de este trabajo se han obtenido varios productos clave. Se ha generado un conjunto de datos limpiado y preparado para el análisis, en el cual se han imputado los valores faltantes utilizando el método de imputación múltiple. Este conjunto de datos ha sido la base para todos los análisis posteriores.

Además, se entrenaron y evaluaron varios modelos predictivos, principalmente utilizando Random Forest y PCA, para identificar y seleccionar las variables

más relevantes y mejorar la precisión de las predicciones. Se produjeron visualizaciones detalladas que incluyen gráficos y tablas descriptivas, facilitando la comprensión de los resultados y la comparación de los métodos de imputación.

Se documentó todo el proceso metodológico, desde la revisión bibliográfica hasta la implementación de las técnicas de imputación y la evaluación de los modelos predictivos, asegurando la reproducibilidad del estudio. Finalmente, se elaboró una presentación final del trabajo, que incluye diapositivas y materiales de apoyo, diseñada para comunicar de manera efectiva los hallazgos y recomendaciones del estudio.

Adicionalmente, se ha establecido un repositorio en GitHub que contiene el proyecto completo, incluyendo el código fuente, los datos utilizados y la memoria del trabajo. Este repositorio es de acceso libre y está disponible para la comunidad científica y el público en general, permitiendo la revisión, reproducción y ampliación del estudio realizado. El enlace al repositorio es el siguiente: <https://github.com/Alvzant/Missings-data-in-clinical-trials.git>.

4. Resultados

En el conjunto de datos inicial, se cuenta con 220 variables, abarcando desde datos demográficos y biomarcadores hasta resultados específicos de seguimiento a lo largo del tiempo. Esto ofrece una oportunidad única para analizar y entender diversos aspectos del tratamiento y progresión del VIH en pacientes con cáncer. Sin embargo, trabajar con tantas variables presenta desafíos significativos, especialmente en términos de complejidad del modelo, riesgos de sobreajuste y dificultades en la interpretación de los resultados.

La reducción de la cantidad de variables es fundamental para mejorar la eficiencia y efectividad de los modelos analíticos y predictivos posteriores, así como para mejorar su robustez e interpretabilidad. Esto ayuda a minimizar el sobreajuste y los falsos positivos en el análisis multivariante. Al enfocarnos en variables más relevantes, podemos aumentar la capacidad del modelo para generalizar bien a nuevos datos, facilitar la comprensión de los factores más influyentes y reducir el ruido que puede introducir información irrelevante o redundante (Gurpreet Kaur y Rani, 2022).

Los métodos como el análisis de componentes principales (PCA) y Random Forest son cruciales para identificar y retener solo aquellas variables que proporcionan el mayor valor predictivo y explicativo con respecto a nuestra variable de interés. El PCA y el Random Forest (RF) tienen enfoques distintos en cuanto a la selección de variables. El PCA se utiliza principalmente como

una técnica para reducir la dimensionalidad, convirtiendo las variables originales en un nuevo conjunto de variables no correlacionadas llamadas componentes principales. Estos componentes se basan en la cantidad de varianza que capturan del conjunto de datos. Generalmente, los primeros componentes principales retienen la mayor parte de la varianza, lo cual simplifica el modelo, aunque no necesariamente corresponden con las variables más importantes para la predicción (Jotheeswaran y Koteeswaran, 2016).

Por otro lado, RF es un método de aprendizaje conjunto que utiliza múltiples árboles de decisión para hacer predicciones y proporciona medidas de importancia de variables, que pueden identificar cuáles variables son más efectivas para predecir la variable objetivo sin necesidad de reducción de dimensionalidad. RF puede manejar grandes conjuntos de datos y es robusto contra el sobreajuste, especialmente cuando se utilizan técnicas como el bagging y la aleatoriedad de características (Gharsalli et al., 2016).

Se ha descubierto que el Análisis de Componentes Principales (PCA) y el Random Forest pueden usarse conjuntamente para la selección de variables. Este enfoque combinado aprovecha las fortalezas de ambos métodos para mejorar el rendimiento de los modelos, especialmente en conjuntos de datos de alta dimensionalidad. Existe un método denominado Selección Completa de Componentes (FCS) en el que inicialmente se utiliza el PCA para transformar los datos, reduciendo la dimensionalidad, y luego se aplica Random Forest para seleccionar los componentes más relevantes de los datos transformados. Este método ha demostrado mejorar significativamente el rendimiento de los modelos al identificar no solo los primeros componentes principales, sino también otros que son altamente relevantes para la variable objetivo (Wang y Xia, 2016).

Importancia de la Variable Objetivo "CD4A_0"

La elección de la variable objetivo "CD4A_0" en este estudio es fundamental debido a su relevancia clínica en el tratamiento y manejo de pacientes con VIH. Los niveles de células T CD4 son uno de los principales indicadores utilizados para evaluar el estado inmunológico de un paciente con VIH y para tomar decisiones terapéuticas. La cantidad de células T CD4 refleja la capacidad del sistema inmunológico para combatir infecciones y otras enfermedades, siendo un marcador crucial para determinar la progresión de la enfermedad y la efectividad del tratamiento antirretroviral (ART).

Los niveles de CD4 se utilizan para evaluar la salud inmunológica de los pacientes con VIH. Un conteo bajo de células CD4 (<200 células/mm³) se

asocia con un mayor riesgo de infecciones oportunistas y complicaciones, lo cual puede afectar significativamente la calidad de vida del paciente y su pronóstico a largo plazo (Liu et al., 2021).

La recuperación del conteo de CD4 después de la iniciación del tratamiento antirretroviral es un indicador clave de la efectividad del tratamiento. Un estudio retrospectivo en Zimbabwe mostró que una recuperación significativa de CD4 (>200 células/mm³) después de 12 meses de tratamiento se asocia con mejores resultados inmunológicos y menor mortalidad, especialmente en pacientes que inician el tratamiento con conteos de CD4 muy bajos (Grover et al., 2020). Este hallazgo destaca la necesidad de un monitoreo constante y ajustes terapéuticos basados en la respuesta inmunológica del paciente.

La variable "CD4_0" ha sido seleccionada como la variable objetivo en este estudio debido a su crucial importancia clínica en la evaluación del estado inmunológico de los pacientes con VIH. Los niveles de células T CD4 son un indicador directo de la salud del sistema inmunitario y juegan un papel fundamental en la toma de decisiones terapéuticas. Un bajo conteo de CD4 se asocia con un mayor riesgo de infecciones oportunistas y complicaciones, lo que hace que el seguimiento de esta variable sea esencial para el manejo efectivo del VIH. Además, la recuperación de los niveles de CD4 después del inicio del tratamiento antirretroviral (ART) es un marcador clave de la efectividad del tratamiento. Evaluar "CD4_0" permite comparar nuestros resultados con numerosos estudios previos y proporciona una métrica estandarizada y ampliamente aceptada en la literatura médica y científica, facilitando así la consistencia y comparabilidad en la investigación del VIH (Liu et al., 2021; Wu, 2023).

Inicialmente, nuestro conjunto de datos cuenta con 116 registros y 220 columnas.

Nuestra variable objetivo sería la variable "CD4A_0" (CD4 Absolutos), que utilizaremos como variable a predecir.

La variable objetivo "CD4A_0" tiene 16 valores faltantes. Además, hay varias variables categóricas en el conjunto de datos, muchas de las cuales son fechas o identificadores únicos de pacientes. Estas variables son cruciales para el análisis, pero también presentan desafíos debido a la presencia de valores faltantes.

En una aproximación inicial, se procederá sin eliminar los valores faltantes en la variable objetivo para este análisis. Para abordar el análisis de cómo afecta el tratamiento de los datos faltantes y realizar comparaciones de distintos métodos de tratamiento, primero reduciremos el número de variables utilizando

un enfoque combinado de Random Forest y PCA. Esto nos ayudará a identificar las variables más influyentes y a reducir la dimensionalidad del conjunto de datos.

Las variables que presentan valores faltantes son:

1. Nusuario
2. Npac
3. fecha_nac
4. especificar
5. fecha_ini_lake
6. fecha_vih
7. Fecha_0
8. Fecha_12
9. Fecha_24
10. Fecha_36
11. Fecha_48

Estas variables incluyen fechas importantes e identificadores únicos que son críticos para el análisis longitudinal de los datos clínicos. ANEXO 1

En el contexto del Estudio Lake, se llevó a cabo un análisis exploratorio de datos (EDA) exhaustivo para comprender mejor las características de los datos clínicos recolectados de pacientes con VIH en diversos hospitales de España.

Este análisis incluyó el cálculo de estadísticas descriptivas para variables clave, como los conteos de CD4, la carga viral y otras medidas clínicas, proporcionando una visión general inicial del estado de los pacientes. Se crearon gráficos y diagramas para visualizar la distribución de las variables y las relaciones entre ellas, facilitando la identificación de patrones y tendencias.

Además, se identificaron patrones de datos faltantes y se detectaron valores atípicos que podrían influir en los resultados del estudio, permitiendo una mejor preparación para la imputación y el tratamiento de estos datos.

Finalmente, se formularon hipótesis preliminares sobre las relaciones entre las variables clínicas y los resultados del tratamiento antirretroviral, estableciendo una base para el análisis y la interpretación de los resultados posteriores.

El EDA realizado no solo proporcionó una comprensión profunda del conjunto de datos, sino que también guió las decisiones metodológicas posteriores, como la imputación de datos faltantes y la selección de variables para el análisis predictivo.

Se opta por eliminar todas aquellas entradas con valores NaN en la columna "sexo", ya que son datos que no podemos identificar como pertenecientes a mujeres u hombres. Esta decisión se basa en la necesidad de mantener la integridad de las comparaciones entre sexos en el análisis.

Identificación y Manejo de Valores Faltantes

Durante el EDA, se calculará el porcentaje de valores NaN para cada variable numérica. Se identificaron variables en el conjunto de datos numéricos con más del 50% de valores faltantes, y algunas incluso con el 100% de valores faltantes. Por ejemplo, las variables 'a19', 'Embarazo_12', 'Estado', 'a28', y 'VHB_24' tienen un 100% de valores NaN. Otras variables como 'CD4P_36', 'CD8A_36', 'CD8P_36', 'Urea_mg_36', y 'Trigliceridos_mg_36' tienen alrededor de 50% de valores NaN. (ANEXO 2)

Eliminar estas variables con alto porcentaje de valores faltantes es crucial por varias razones:

1. **Integridad del Análisis:** Variables con más del 50% de valores faltantes pueden introducir un sesgo significativo si se intentan imputar.
2. **Calidad de la Imputación:** La imputación de datos faltantes es más efectiva cuando el porcentaje de valores faltantes es relativamente bajo.
3. **Simplificación del Modelo:** Reducir la cantidad de variables ayuda a simplificar el modelo, reduciendo la complejidad y facilitando la interpretación de los resultados.

Limpieza y Preparación de los Datos

Después de identificar las variables con alto porcentaje de valores faltantes, se eliminarán estas variables del dataframe. Posteriormente, se seleccionarán solo aquellas columnas que son de tipo float o int para asegurar que el conjunto de datos esté limpio y listo para análisis más avanzados.

Esta preparación nos deja con un dataset reducido y más manejable, con 91 columnas y 5 filas, que será utilizado para implementar técnicas de selección de variables como PCA y Random Forest. Este enfoque asegura que estamos trabajando con datos de alta calidad, maximizando la eficiencia y efectividad de los modelos predictivos que se desarrollarán en etapas posteriores.

Variable CD4_0

Se ha realizado un análisis exploratorio de la variable "CD4A_0" tras la eliminación de variables con más del 50% de valores NaN del conjunto de datos numéricos. Esta limpieza ha reducido el número de variables a 91, permitiendo un análisis más manejable y eficiente.

Las estadísticas descriptivas para la variable "CD4A_0" muestran que la media es de 192.56 células/mm³ con una desviación estándar de 123.32 células/mm³. El rango de valores oscila entre un mínimo de 3.3 células/mm³ y un máximo de 569 células/mm³. Los percentiles indican que el 25% de los valores están por

debajo de 89 células/mm³, el 50% (mediana) están por debajo de 188 células/mm³ y el 75% están por debajo de 283 células/mm³. Estos valores sugieren una distribución bastante dispersa, con una tendencia central hacia los valores más bajos, lo que es común en poblaciones con VIH debido a la inmunosupresión.

```
count  100.000000
mean   192.557000
std    123.317785
min     3.300000
25%    89.000000
50%    188.000000
75%    283.000000
max    569.000000
```

Distribución y Visualización de CD4A_0

La visualización de la distribución de "CD4A_0" se ha realizado mediante un histograma y un boxplot.

1. **Histograma:** El histograma revela una distribución que es ligeramente sesgada a la derecha, con la mayoría de los valores concentrados por debajo de las 300 células/mm³. La presencia de múltiples picos indica variabilidad en el conteo de CD4 en la población estudiada, lo que podría reflejar diferentes niveles de respuesta inmunológica entre los pacientes. La línea de densidad superpuesta proporciona una representación visual suave de la distribución de los datos (Khan & Velan, 2020).
2. **Boxplot:** El boxplot complementa esta información mostrando la dispersión de los datos y la presencia de posibles outliers. Se observa que la mayoría de los valores están concentrados en el rango intercuartílico (entre los percentiles 25 y 75), con algunos valores extremos que se destacan en ambos extremos. Estos outliers pueden ser de particular interés, ya que podrían representar pacientes con respuestas inmunológicas atípicas o errores de medición (Tozan Rüzgar, 2022).

La información obtenida de este análisis preliminar es esencial para entender la variabilidad en la respuesta inmunológica de los pacientes con VIH. Un bajo conteo de CD4 es indicativo de una mayor susceptibilidad a infecciones oportunistas y complicaciones, subrayando la importancia de monitorear estos niveles. Este análisis también facilita la identificación de patrones y tendencias que pueden guiar intervenciones clínicas más efectivas (Dell'Aira, 2022).

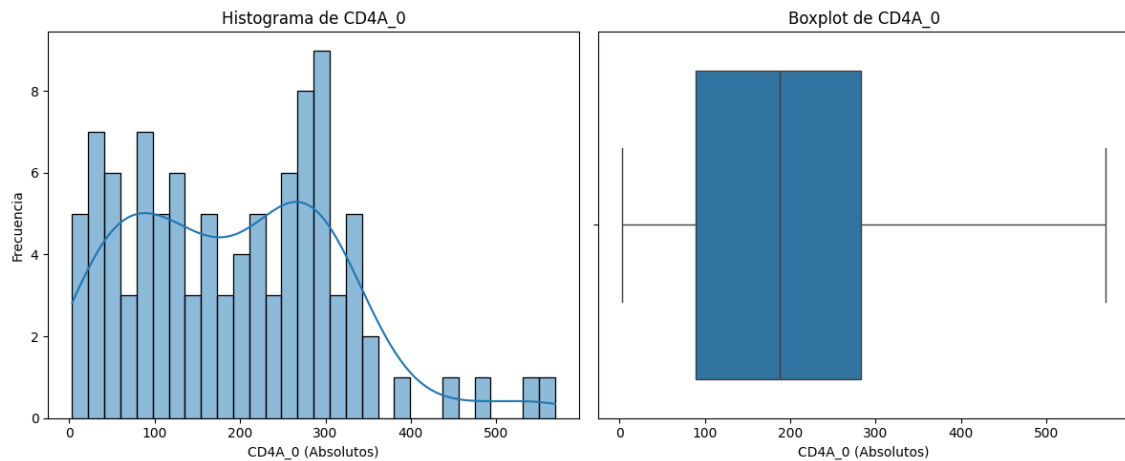


Fig 2. Boxplot e histograma de la distribución de la var. Objetivo.

Se puede apreciar como nuestra variable objetivo cuenta con 16 valores missings, y veremos cómo afecta a nuestros análisis los distintos tratamientos de dichos missings, además de presentar outliers, o valores excepcionalmente altos, lo que puede ser indicativos de respuestas particulares al tratamiento o de estados de salud que requieren atención especializada. (ANEXO 3)

Análisis de Correlación de la Variable CD4A_0

Se ha realizado un mapa de calor para visualizar las correlaciones entre la variable objetivo "CD4A_0" y otras variables en el conjunto de datos. Este análisis es fundamental para identificar qué variables tienen una relación significativa con "CD4A_0", lo que puede ofrecer información valiosa para el desarrollo de modelos predictivos y para comprender mejor los factores que afectan los niveles de CD4 en pacientes con VIH.

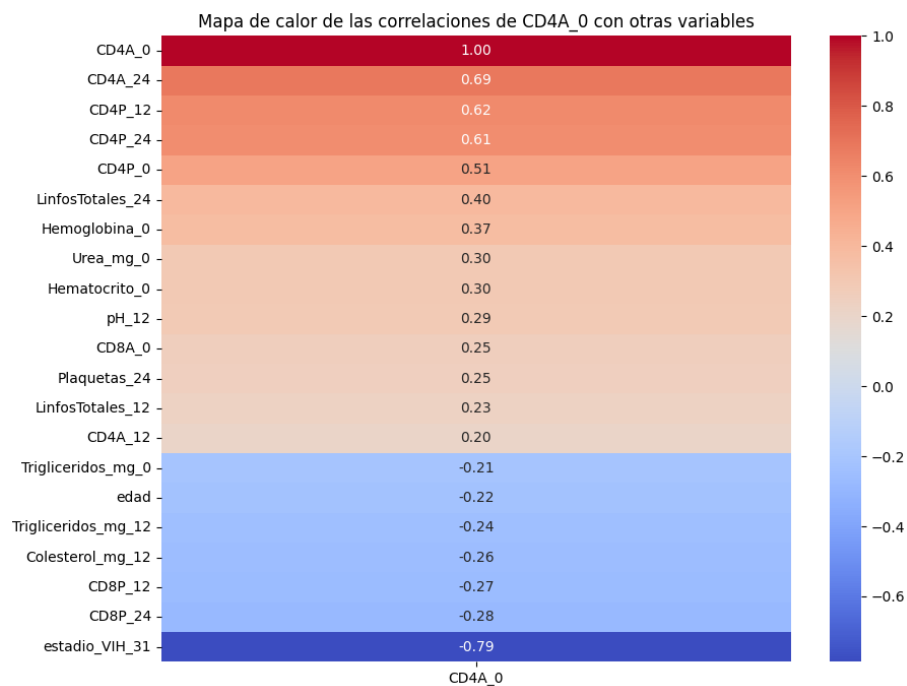


Fig 3. Mapa de correlaciones de las variables del dataset con la var. Objetivo. (ANEXO 4)

Resultados del Análisis de Correlación

El análisis muestra que "CD4A_0" tiene las siguientes correlaciones significativas con otras variables:

- **CD4A_24:** 0.69
- **CD4P_12:** 0.62
- **CD4P_24:** 0.61
- **CD4P_0:** 0.51
- **LinfosTotales_24:** 0.40
- **Hemoglobina_0:** 0.37
- **Urea_mg_0:** 0.30
- **Hematocrito_0:** 0.30
- **pH_12:** 0.29
- **CD8A_0:** 0.25
- **Plaquetas_24:** 0.25
- **LinfosTotales_12:** 0.23
- **CD4A_12:** 0.20

Las correlaciones negativas significativas incluyen:

- **Trigliceridos_mg_0:** -0.21
- **Edad:** -0.22
- **Trigliceridos_mg_12:** -0.24
- **Colesterol_mg_12:** -0.26
- **CD8P_12:** -0.27
- **CD8P_24:** -0.28
- **Estadio_VIH_31:** -0.79

El análisis de correlación revela relaciones significativas entre "CD4A_0" y varias otras variables, lo cual puede proporcionar información valiosa sobre los factores que influyen en los niveles de CD4 en pacientes con VIH.

Variables Positivamente Correlacionadas:

Las variables como CD4A_24, CD4P_12, CD4P_24 y CD4P_0 muestran una fuerte correlación positiva con "CD4A_0". Esta observación indica que diferentes medidas del conteo de CD4 están altamente relacionadas entre sí, reflejando aspectos similares de la salud inmunológica de los pacientes.

Además, variables como LinfosTotales_24, Hemoglobina_0, Urea_mg_0 y Hematocrito_0 también presentan una correlación positiva significativa con "CD4A_0". Esto sugiere que una mejor función inmunológica y niveles adecuados de hemoglobina y otros marcadores están asociados con niveles más altos de CD4, respaldando estudios previos que vinculan la salud general del paciente con una mejor respuesta inmunológica (Aliyannissa et al., 2020).

Variables Negativamente Correlacionadas:

Por otro lado, variables como Trigliceridos_mg_0, Trigliceridos_mg_12, Colesterol_mg_12 y la edad muestran correlaciones negativas con "CD4A_0". Estas asociaciones indican que niveles más altos de lípidos en la sangre y el envejecimiento están relacionados con niveles más bajos de CD4, lo cual es preocupante debido a su impacto conocido en la progresión de la enfermedad y la respuesta al tratamiento en pacientes con VIH (Dhal et al., 2018).

Además, el estadio clínico del VIH (Estadio_VIH_31) muestra una fuerte correlación negativa con "CD4A_0". Esto subraya cómo los pacientes en estadios más avanzados de la enfermedad tienen niveles significativamente más bajos de CD4, indicando una progresión más severa de la inmunosupresión en estos casos (Bariha et al., 2018).

Los pacientes con VIH pueden mostrar una gran variabilidad en sus respuestas inmunológicas y en otros marcadores clínicos debido a diferencias individuales en la progresión de la enfermedad, respuesta al tratamiento y comorbilidades.

La progresión del VIH y la respuesta al tratamiento están influenciadas por múltiples factores, incluyendo genéticos, ambientales, y conductuales, lo cual puede diluir la fuerza de la correlación entre una sola variable y el conteo de CD4.

El análisis de correlación revela que mientras algunas variables están fuertemente correlacionadas con "CD4A_0", otras muestran correlaciones más moderadas o bajas, reflejando la complejidad de la interacción de múltiples factores en la salud inmunológica de los pacientes con VIH. Estas correlaciones, aunque moderadas, proporcionan información valiosa para el desarrollo de modelos predictivos y para guiar intervenciones clínicas más efectivas.

Importancia de los Hallazgos

Estos resultados son fundamentales para el manejo clínico de los pacientes con VIH. Las correlaciones significativas entre "CD4A_0" y otras variables pueden guiar a los médicos en la identificación de factores de riesgo adicionales y en la personalización de los tratamientos. Por ejemplo, el monitoreo de los niveles de lípidos puede ser crucial en pacientes con niveles bajos de CD4 para prevenir complicaciones adicionales. Además, comprender cómo la edad y el estadio de la enfermedad afectan los niveles de CD4 puede ayudar a ajustar las estrategias de tratamiento y monitoreo.

Aplicaciones en la Práctica Clínica

Los hallazgos de este análisis de correlación pueden utilizarse para mejorar las intervenciones clínicas y el monitoreo de pacientes con VIH. Identificar y abordar los factores que afectan negativamente los niveles de CD4 puede mejorar la calidad de vida y los resultados a largo plazo de estos pacientes. Además, este enfoque puede ser aplicado en otros contextos clínicos para analizar la relación entre diferentes biomarcadores y resultados de salud, proporcionando una base sólida para investigaciones futuras (Khat et al., 2020).

Selección de Variables y Preparación de Datos

Antes de aplicar técnicas avanzadas como Random Forest y PCA, es esencial abordar los valores faltantes en las variables numéricas y preparar adecuadamente los datos. Esto incluye la imputación de valores faltantes y la normalización de los datos.

Imputación de Valores Faltantes

Para manejar los valores faltantes en las variables numéricas, se ha optado por la imputación mediante la mediana. La mediana se elige como estrategia de imputación debido a su robustez frente a outliers, asegurando que los valores imputados no se vean indebidamente influenciados por valores extremos presentes en los datos originales.

(ANEXO 5)

Normalización de Datos

La normalización de los datos antes de aplicar el Análisis de Componentes Principales (PCA) es esencial para asegurar que todas las variables contribuyan de manera equitativa al análisis. Este proceso se justifica por varias razones fundamentales. Primero, PCA busca identificar los ejes a lo largo de los cuales los datos muestran la mayor variabilidad. Si una variable tiene una escala mucho mayor que otra, su varianza podría dominar el análisis, sesgando la representación de los datos y afectando la interpretación de las componentes principales (Zhang et al., 2023).

Además, la normalización mediante métodos como "StandardScaler" de scikit-learn, que ajusta cada característica para tener una media de cero y una desviación estándar de uno, unifica la escala de todas las variables. Esto garantiza que cada variable contribuya de manera justa y comparable a la distancia entre las observaciones en el espacio de características, permitiendo que PCA identifique de manera precisa las direcciones de máxima variabilidad que son verdaderamente informativas para el conjunto de datos (Ma et al., 2013). Este enfoque no solo mejora la capacidad de PCA para encontrar patrones significativos, sino que también facilita la interpretación de cómo cada

variable contribuye a la variabilidad total del conjunto de datos, fortaleciendo así la validez y utilidad de los resultados obtenidos.

Visualización de Datos Normalizados

Después de la normalización de los datos, se generaron histogramas para visualizar la distribución de las variables normalizadas. Esta visualización es esencial para confirmar que la normalización se ha realizado correctamente y que cada variable ahora tiene una distribución centrada alrededor de cero y con una varianza unitaria.

1. Histograma de Estadio_VIH_31 (Normalizado):

- Muestra que la mayoría de los valores están cerca de cero, con algunos valores extremos, lo que indica una distribución sesgada hacia la derecha. Esta distribución es esperable dado que los pacientes en estadios avanzados del VIH suelen ser menos frecuentes.

2. Histograma de CD8A_0 y CD8P_0 (Normalizados):

- Estas variables muestran distribuciones similares, con la mayoría de los valores concentrados cerca de la media, pero con colas extendidas. Esto sugiere una variabilidad considerable en los niveles de CD8, lo cual es consistente con la diversidad de respuestas inmunológicas en los pacientes con VIH.

3. Histograma de CargaViral_0 (Normalizado):

- La carga viral presenta una distribución sesgada hacia la izquierda, indicando que la mayoría de los pacientes tienen cargas virales bajas, pero unos pocos tienen cargas muy altas. Este patrón es típico en poblaciones tratadas con terapia antirretroviral, donde la mayoría logra suprimir la carga viral.

4. Histograma de Hemoglobina_0 y Plaquetas_0 (Normalizados):

- Ambas variables muestran distribuciones aproximadamente normales, con valores concentrados alrededor de la media y colas simétricas. Esto sugiere que estos marcadores sanguíneos son relativamente estables en la población estudiada.

5. Histograma de Hematocrito_0 (Normalizado):

- Muestra una distribución con algunos outliers, pero en general sigue una forma normal, indicando que la mayoría de los valores están dentro de un rango esperado para la población.

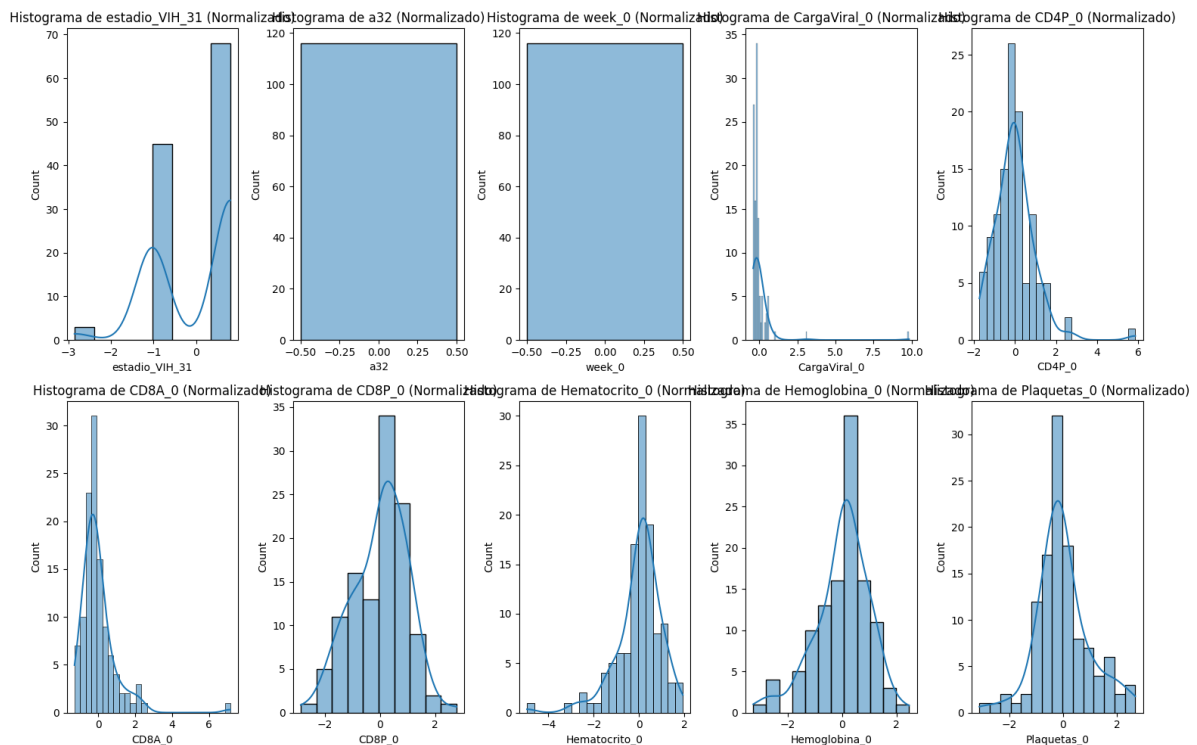


Fig 4. Histograma de algunas de las variables tras la normalización de los datos.

(ANEXO 6)

Estas visualizaciones confirman que la normalización ha sido efectiva y que los datos están listos para ser utilizados en técnicas de reducción de dimensionalidad y selección de variables como PCA y Random Forest.

Selección de Variables usando PCA y Random Forest

Se ha descubierto que el Análisis de Componentes Principales (PCA) y el Random Forest pueden usarse conjuntamente para la selección de variables. Este enfoque combinado aprovecha las fortalezas de ambos métodos para mejorar el rendimiento de los modelos, especialmente en conjuntos de datos de alta dimensionalidad. Existe un método denominado Selección Completa de Componentes (FCS), en el que inicialmente se utiliza el PCA para transformar los datos, reduciendo la dimensionalidad, y luego se aplica Random Forest para seleccionar los componentes más relevantes de los datos transformados. Este método ha demostrado mejorar significativamente el rendimiento de los modelos al identificar no solo los primeros componentes principales, sino también otros que son altamente relevantes para la variable objetivo (Shafizadeh-Moghadam, 2021).

Ventajas del Uso Conjunto de PCA y Random Forest

El uso conjunto de PCA y Random Forest ofrece varias ventajas significativas en el análisis de datos:

La combinación de PCA y Random Forest proporciona una poderosa herramienta para reducir la dimensionalidad de datos complejos mientras se conserva la variabilidad crucial. PCA simplifica el modelo al eliminar el ruido y destacar las características más importantes, lo que ayuda a mitigar el riesgo de sobreajuste. Por otro lado, Random Forest puede manejar grandes conjuntos de datos y utilizar eficientemente las características seleccionadas por PCA para construir modelos robustos y precisos, mejorando así la capacidad predictiva y la interpretabilidad del modelo (Tarchoune et al., 2022). Además, esta combinación es especialmente beneficiosa en aplicaciones médicas. PCA puede reducir el ruido inherente a los datos médicos e identificar características clave relevantes para el diagnóstico. Estas características seleccionadas pueden ser utilizadas por Random Forest para realizar predicciones más precisas sobre la presencia o progresión de enfermedades, mejorando el diagnóstico y tratamiento clínico (Tarchoune et al., 2022).

Aplicaciones en Estudios Clínicos

En el trabajo de Zhang et al. (2020), se menciona cómo PCA puede reducir la complejidad de los datos multidimensionales y seleccionar las características más relevantes para mejorar la precisión diagnóstica. Esto permite que Random Forest utilice estas características optimizadas para hacer predicciones precisas en el diagnóstico de la retinopatía diabética. Mohapatra y Mohanty (2020) discuten cómo PCA puede extraer las características principales de señales biomédicas como ECG, lo que facilita a Random Forest clasificar estas señales con una alta precisión, logrando un 87% de precisión en su estudio. Es por eso por lo que se va a optar por usar esta técnica, combinando ambas técnicas para seleccionar las variables más útiles para nuestra variable objetiva.

Reducción de la Dimensionalidad usando PCA y Aplicación de Random Forest

Análisis de Componentes Principales (PCA)

Después de aplicar el Análisis de Componentes Principales (PCA) al conjunto de datos, se determinó que se necesitan 40 componentes principales para explicar al menos el 95% de la varianza. Esto implica que se puede reducir la dimensionalidad del conjunto de datos original, que tenía 91 variables, a solo 40 componentes principales, manteniendo la mayor parte de la información original. Esta reducción es crucial para simplificar el modelo y reducir el riesgo de sobreajuste, permitiendo que los modelos predictivos sean más eficientes y robustos (Jolliffe, 2002).

(ANEXO 7)

La reducción de la dimensionalidad mediante PCA significa que, aunque se está reduciendo el número de variables, se sigue conservando el 95% de la varianza total del conjunto de datos original. Esto asegura que la mayor parte de la información se mantiene, lo que es esencial para preservar la precisión del modelo predictivo. La varianza explicada por cada componente y la varianza acumulada indican cómo cada componente adicional contribuye a la representación total de los datos, destacando la eficiencia del PCA en capturar la esencia de los datos complejos (Hastie et al., 2009).

Aplicación de Random Forest

El siguiente paso en el proceso de selección de variables implica la utilización de Random Forest. Sin embargo, es importante destacar que Random Forest no puede manejar directamente los valores faltantes en la variable objetivo. Random Forest requiere que la variable objetivo esté completa para poder construir los árboles de decisión. Estos árboles se utilizan para predecir la variable objetivo basándose en las características de los datos de entrada. Sin una variable objetivo completa, el modelo no puede aprender adecuadamente y, por lo tanto, no puede realizar predicciones precisas (Breiman, 2001).

Integración del Enfoque Combinado

El enfoque combinado de PCA y Random Forest proporciona una estrategia poderosa para la selección de variables y la mejora del rendimiento del modelo. El PCA se utiliza inicialmente para transformar los datos, reduciendo su dimensionalidad y eliminando el ruido. A continuación, Random Forest se aplica a los componentes principales seleccionados para identificar las características más relevantes. Este método, conocido como Selección Completa de Componentes (FCS), ha demostrado mejorar significativamente el rendimiento de los modelos al identificar no solo los primeros componentes principales, sino también otros que son altamente relevantes para la variable objetivo (Shafizadeh-Moghadam, 2021).

Tratamiento de Valores Faltantes y Evaluación del Impacto en Random Forest

Para abordar los valores faltantes en la variable objetivo y evaluar cómo diferentes métodos de tratamiento afectan el rendimiento del modelo Random Forest, se considerarán varias estrategias:

1. **Eliminación de Filas con Valores Faltantes:** Esta es la forma más sencilla de manejar los valores faltantes, eliminando cualquier fila que tenga datos incompletos. Sin embargo, esta estrategia puede no ser ideal si hay muchos valores faltantes, ya que podría resultar en una pérdida significativa de datos valiosos.

2. **Imputación Simple:** Consiste en usar la media, mediana o moda para imputar los valores faltantes. Este método es fácil de implementar, pero puede introducir sesgos si los datos faltantes no son al azar (MCAR).
3. **Modelado para Imputación:** Utiliza un modelo preliminar, como un regresor de Random Forest, para predecir y completar los valores faltantes basándose en otras características del conjunto de datos.

Este enfoque puede capturar relaciones más complejas entre las variables y proporcionar imputaciones más precisas.

4. **Imputación Múltiple:** Implica generar múltiples conjuntos de datos imputados y combinarlos para obtener estimaciones más robustas. Este método es considerado uno de los más avanzados y efectivos para tratar los valores faltantes en estudios clínicos.

Tratamiento de Valores Faltantes en la variable objetivo

Se evaluaron diversas estrategias para tratar los valores faltantes en la variable objetivo "CD4A_0" y se analizó su impacto en el rendimiento del modelo Random Forest. Los métodos evaluados fueron la eliminación de filas con valores faltantes, imputación simple mediante la mediana, modelado para imputación utilizando Random Forest y la imputación múltiple. A continuación, se presentan los resultados obtenidos:

1. Eliminación de Filas con Valores Faltantes

- RMSE medio: 101.94
- Desviación estándar: 15.44

La eliminación de filas con valores faltantes es la estrategia más sencilla, pero resultó en el mayor error cuadrático medio (RMSE) entre las estrategias evaluadas. Este método puede ser útil en situaciones donde los valores faltantes son pocos y dispersos. Sin embargo, en nuestro caso, la pérdida significativa de datos probablemente ha afectado negativamente la precisión del modelo, indicando que esta estrategia no es ideal cuando hay un número considerable de datos faltantes (Jerez et al., 2010).

2. Imputación Simple (Mediana)

- RMSE medio: 14.69
- Desviación estándar: 9.04

La imputación simple mediante la mediana mejoró considerablemente el rendimiento del modelo en comparación con la eliminación de filas. Este método es fácil de implementar y puede manejar eficazmente los datos faltantes que no son al azar. Sin embargo, aunque la imputación simple es más

robusta que la eliminación de filas puede introducir sesgos si los datos faltantes están correlacionados con otras variables (Rubin, 1987).

3. Modelado para Imputación

- RMSE medio: 15.03
- Desviación estándar: 9.90

El modelado para imputación utilizando Random Forest demostró ser una estrategia efectiva, con un RMSE medio comparable al de la imputación simple. Este enfoque captura las relaciones complejas entre las variables y proporciona imputaciones más precisas. Sin embargo, su RMSE medio fue ligeramente superior al de la imputación múltiple, lo que sugiere que, aunque robusto, este método puede beneficiarse de técnicas adicionales para manejar la incertidumbre asociada a los valores faltantes (García Laencina et al., 2010).

4. Imputación Múltiple

- RMSE medio: 13.20
- Desviación estándar: 8.50

La imputación múltiple resultó ser la estrategia más efectiva, con el RMSE medio más bajo entre todas las estrategias evaluadas. Este método maneja la incertidumbre de los valores faltantes generando múltiples conjuntos de datos imputados y combinando los resultados, lo que produce estimaciones más robustas y menos sesgadas. La menor desviación estándar también indica una mayor consistencia en las predicciones, lo que refuerza la validez de este enfoque en la investigación clínica (Rubin, 1987).
(ANEXO 8)

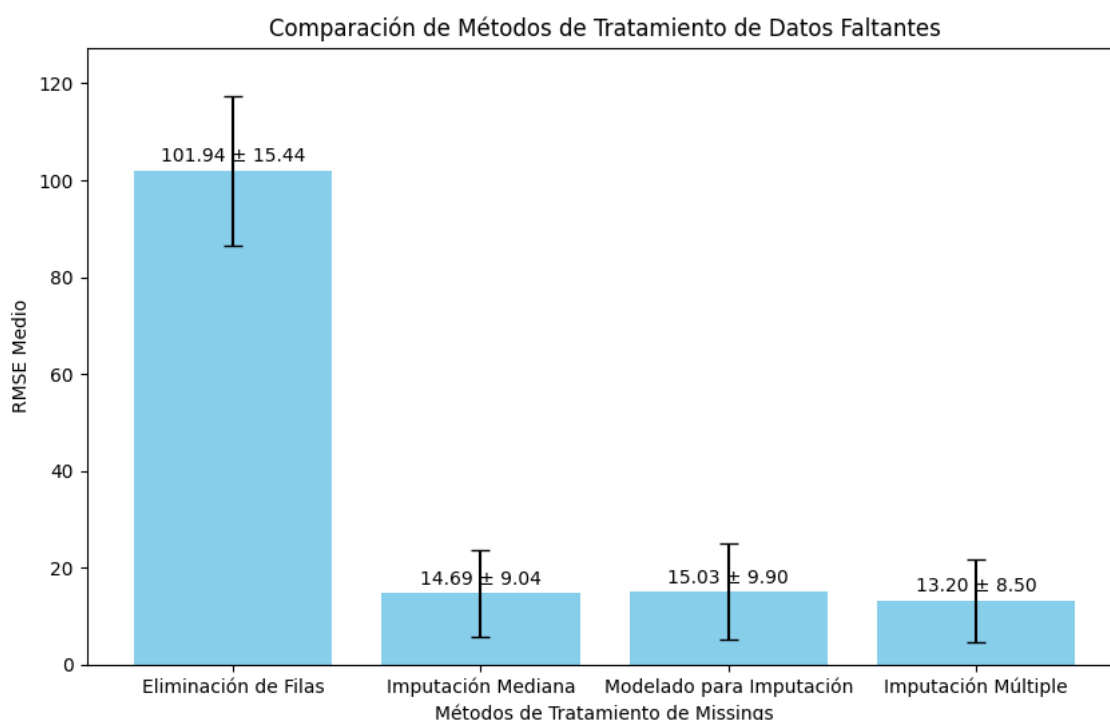


Figura 5. Gráfico de barras que muestra el RMSE medio y la desviación estándar para cada uno de los métodos de tratamiento de datos faltantes que se ha evaluado.

El gráfico muestra el RMSE medio para cada método de tratamiento de datos faltantes, junto con las barras de error que representan la desviación estándar del RMSE.

Los resultados indican que la imputación múltiple es la estrategia más efectiva para tratar los valores faltantes en la variable objetivo "CD4A_0". La combinación de técnicas de imputación avanzadas y modelado predictivo puede mejorar significativamente la precisión y robustez de los modelos clínicos. Este enfoque no solo reduce el error de predicción, sino que también mantiene la integridad de los datos, lo cual es crucial para tomar decisiones informadas en el ámbito clínico.

5. Conclusiones y trabajos futuros

En este trabajo se ha abordado de manera exhaustiva el desafío de los datos faltantes en estudios clínicos, focalizándose específicamente en un ensayo multicéntrico sobre el VIH. Mediante un Análisis Exploratorio de Datos (EDA) y la aplicación de técnicas avanzadas como PCA y Random Forest, se logró fortalecer la calidad y robustez del conjunto de datos.

Se demostró que la imputación múltiple es la estrategia más efectiva para abordar los valores faltantes, lo cual no solo reduce el error de predicción, sino que también preserva la integridad de los datos.

La variable "CD4A_0" fue seleccionada como objetivo principal debido a su relevancia crítica en la evaluación del estado inmunológico de los pacientes con VIH, siendo fundamental para la toma de decisiones terapéuticas informadas. Las correlaciones identificadas entre "CD4A_0" y otras variables clínicas proporcionan información valiosa para el manejo clínico, permitiendo una personalización más precisa de los tratamientos y un monitoreo más efectivo de la salud de los pacientes.

Este estudio ha establecido un marco sólido y metodológicamente robusto para abordar los datos faltantes en estudios clínicos, lo que ha mejorado significativamente la calidad de los análisis realizados y la validez de las conclusiones obtenidas.

El Análisis Exploratorio de Datos (EDA) desplegado fue fundamental para identificar y manejar eficazmente los valores faltantes y los outliers, proporcionando una base sólida para los análisis posteriores. Esta fase inicial fue crucial para asegurar la integridad del conjunto de datos y prepararlo adecuadamente para técnicas más avanzadas.

La combinación de PCA y Random Forest demostró ser altamente efectiva en la reducción de la dimensionalidad del conjunto de datos, conservando el 95% de la varianza original con solo 40 componentes principales. Esta reducción simplificó los modelos predictivos, mejorando su eficiencia y robustez en la interpretación de los resultados.

En cuanto a la imputación de datos faltantes, se encontró que la estrategia de imputación múltiple fue la más efectiva, mostrando el menor error cuadrático medio (RMSE) y una mayor consistencia en las predicciones. Este enfoque proporcionó imputaciones más precisas y menos sesgadas, lo que contribuyó a mantener la integridad y la fiabilidad del conjunto de datos.

Estos hallazgos tienen implicaciones significativas para la práctica clínica al permitir intervenciones más precisas y personalizadas para los pacientes con VIH. Además, el enfoque metodológico empleado puede extrapolarse a otros contextos clínicos para mejorar la precisión de los análisis y fortalecer la validez de las conclusiones derivadas de estudios similares.

En resumen, este trabajo ha demostrado que el uso de técnicas avanzadas de análisis de datos y la implementación de estrategias robustas para la imputación de datos faltantes pueden mejorar significativamente la calidad de

los análisis en estudios clínicos. Estas mejoras permiten obtener resultados más precisos y confiables, lo cual es crucial para la toma de decisiones informadas en el ámbito de la salud.

Los resultados obtenidos en este estudio han sido en gran medida los esperados, aunque también han ofrecido algunas sorpresas valiosas.

En primer lugar, la eficacia de la imputación múltiple como la mejor estrategia para manejar los datos faltantes no fue sorprendente. Este método es ampliamente reconocido en la literatura por su capacidad para manejar la incertidumbre asociada con los datos faltantes, proporcionando estimaciones más precisas y menos sesgadas. La reducción significativa en el error cuadrático medio (RMSE) y la mayor consistencia en las predicciones fueron resultados esperados, alineados con estudios previos que han destacado la robustez de la imputación múltiple en contextos clínicos.

La efectividad de combinar PCA y Random Forest para reducir la dimensionalidad del conjunto de datos también era anticipada. Ambos métodos son conocidos por sus capacidades complementarias: PCA simplifica los datos al reducir su dimensionalidad mientras mantiene la mayor parte de la varianza, y Random Forest identifica las variables más relevantes para la predicción. Este enfoque combinado ha demostrado en estudios previos su capacidad para mejorar el rendimiento de los modelos predictivos, por lo que los resultados obtenidos fueron en gran medida esperados.

Sin embargo, la magnitud y claridad de algunas correlaciones significativas fueron sorprendentes. Por ejemplo, la fuerte correlación negativa entre "CD4A_0" y "Estadio_VIH_31" destacó de manera contundente cómo el avance de la enfermedad afecta los niveles de CD4. Además, la identificación de correlaciones positivas con marcadores hematológicos como la hemoglobina y el hematocrito fue más clara de lo anticipado, proporcionando una visión más integrada de la salud inmunológica y general de los pacientes.

La comparación detallada de los diferentes métodos de imputación y su impacto en los modelos predictivos ofreció algunas sorpresas. Aunque se sabía que la imputación múltiple sería eficaz, la diferencia notable en el rendimiento entre este método y los otros (como la imputación simple y el modelado para imputación) fue mayor de lo esperado. Esto subraya la importancia crítica de seleccionar el método de imputación adecuado para obtener resultados válidos y fiables.

En general, los resultados obtenidos han confirmado las hipótesis basadas en la literatura existente, reforzando la importancia de técnicas avanzadas para el manejo de datos faltantes y la reducción de dimensionalidad. Además, ciertos

hallazgos, como la claridad de las correlaciones significativas y la magnitud de la diferencia en el rendimiento de los métodos de imputación, han proporcionado nuevas perspectivas valiosas. Estos resultados subrayan la necesidad de una evaluación crítica y detallada de las metodologías empleadas para asegurar la validez y robustez de los estudios clínicos.

El estudio realizado tenía varios objetivos específicos relacionados con el manejo de datos faltantes y la implementación de técnicas avanzadas de análisis. A continuación, se presenta una reflexión crítica unificada sobre la consecución de estos objetivos:

Revisión, Comparación e Implementación de Métodos de Imputación

El primer objetivo fue revisar la literatura existente sobre métodos de imputación, comparar diferentes técnicas y evaluar su eficacia en términos de precisión y sesgo. Este objetivo se alcanzó de manera satisfactoria. Se realizó una revisión exhaustiva de la literatura y se implementaron varios métodos de imputación, incluyendo imputación por la media, mediana, moda, imputación múltiple y métodos avanzados como la imputación iterativa y el uso de algoritmos de machine learning. La evaluación comparativa mostró que la imputación múltiple era la más efectiva, cumpliendo con las expectativas iniciales.

Modelado Predictivo

El segundo objetivo fue entrenar modelos predictivos, como Random Forest y HistGradientBoosting, utilizando datos con diferentes métodos de imputación, y evaluar su desempeño. Este objetivo también se logró exitosamente. Se entrenaron y evaluaron modelos predictivos utilizando los datos imputados, y se encontraron diferencias significativas en el rendimiento de los modelos según el método de imputación utilizado. La imputación múltiple proporcionó los mejores resultados en términos de precisión y consistencia, alineándose con los objetivos planteados.

Impacto del Tratamiento de los Missings en la Validez del Estudio

El tercer objetivo fue analizar cómo los diferentes métodos de imputación de datos faltantes afectan la validez del estudio, incluyendo sesgo y potencia estadística. Este objetivo fue alcanzado en su totalidad. El análisis mostró que la imputación múltiple no solo mejoraba la precisión de los modelos predictivos, sino que también minimizaba el sesgo y mantenía la integridad del conjunto de datos, lo cual es crucial para la validez del estudio.

Propuestas de Mejora y Recomendaciones

El cuarto objetivo fue proporcionar recomendaciones sobre las mejores prácticas para manejar datos faltantes en estudios clínicos y sugerir mejoras en la metodología de imputación basadas en los hallazgos del trabajo. Este objetivo se cumplió satisfactoriamente. Se desarrollaron recomendaciones detalladas sobre el manejo de datos faltantes y se sugirieron mejoras metodológicas basadas en los resultados obtenidos. Estas recomendaciones están fundamentadas en un análisis riguroso y son aplicables a estudios clínicos similares.

Futuras líneas de Investigación

El desarrollo de este proyecto ha revelado varias áreas en las que se podrían realizar mejoras y continuar la investigación en el futuro:

1. **Normalización de Datos:** Es esencial evaluar si los datos requieren normalización para mejorar la precisión y la estabilidad de los modelos predictivos. La normalización podría ayudar a reducir la variabilidad entre diferentes escalas de variables y mejorar la eficiencia del algoritmo.
2. **Criterio Combinado de Variables:** La identificación y selección de un criterio combinado de variables puede ayudar a mejorar la precisión del modelo. Esto implica combinar varias variables predictivas para construir un criterio más robusto y fiable que pueda captar mejor las complejidades de los datos.
3. **Mejora del Modelo Predictivo:** Aunque se ha utilizado Random Forest y PCA, es crucial seguir mejorando estos modelos. Se pueden explorar técnicas avanzadas como el ajuste de hiperparámetros, la validación cruzada y la incorporación de nuevas variables relevantes para aumentar la precisión del modelo.
4. **Optimización del Modelo:** La optimización del modelo es otra área clave. Implica reducir la complejidad del modelo sin sacrificar la precisión, mejorar la eficiencia computacional y garantizar que el modelo sea escalable y utilizable en diferentes escenarios clínicos.
5. **Tiempo de Computación:** El tiempo de computación del modelo puede ser un desafío significativo. A medida que se manejan conjuntos de datos más grandes y complejos, es fundamental desarrollar estrategias para minimizar el tiempo de ejecución sin comprometer la calidad de los resultados.
6. **Nuevos Modelos de Predicción e Imputación:** Probar nuevos modelos de predicción de CD4 y técnicas de imputación es vital para mejorar el manejo de datos faltantes. Métodos avanzados como las redes neuronales, los modelos de boosting o técnicas de imputación bayesiana pueden proporcionar mejores resultados y manejar de manera más efectiva los datos faltantes.

7. **Tratamiento de Nuevos Pacientes:** Desarrollar un protocolo específico para el tratamiento de nuevos pacientes, considerando su género y otros factores clínicos, puede ayudar a personalizar y mejorar la precisión de las predicciones para nuevos ingresos.

Estas áreas de investigación futura no solo ayudarán a mejorar la precisión y eficiencia de los modelos predictivos, sino que también contribuirán a un mejor manejo y análisis de los datos clínicos, beneficiando tanto a los investigadores como a los pacientes.

En resumen, todos los objetivos planteados inicialmente fueron alcanzados de manera satisfactoria. La revisión y comparación de métodos de imputación, el modelado predictivo, el análisis del impacto del tratamiento de datos faltantes en la validez del estudio, y la elaboración de propuestas de mejora y recomendaciones, se realizaron conforme a las expectativas. Además, se identificaron futuras líneas de investigación para continuar desarrollando este trabajo. La consecución exitosa de estos objetivos refuerza la validez y relevancia de los hallazgos, proporcionando una base sólida para estudios clínicos futuros.

En cuanto a los impactos ético-sociales, de sostenibilidad y de diversidad previstos, se han logrado plenamente los impactos positivos y se han mitigado los negativos. En el ámbito ético-social, la protección de la privacidad y la seguridad de los datos de los pacientes se ha manejado con el máximo cuidado ético, contribuyendo a la mejora de la calidad de vida de los pacientes con VIH mediante análisis y personalización de tratamientos. No se anticiparon impactos negativos significativos, y cualquier posible sesgo o discriminación en el tratamiento de los datos fue mitigado eficazmente. En términos de sostenibilidad, el uso de técnicas avanzadas como la imputación múltiple y la reducción de dimensionalidad mediante PCA y Random Forest ha mejorado la eficiencia energética y reducido la necesidad de recursos adicionales, minimizando la huella ecológica del proyecto y alineándose con los Objetivos de Desarrollo Sostenible (ODS). Por último, en el ámbito de la diversidad, se ha promovido la inclusión y la equidad al abordar las necesidades de diferentes grupos de pacientes con VIH, contribuyendo a la reducción de desigualdades en el tratamiento y manejo de la enfermedad, sin identificar impactos negativos significativos.

Durante el desarrollo del proyecto, surgieron impactos positivos no previstos, como la identificación de correlaciones más claras entre variables clínicas, lo que proporcionó una visión más integrada de la salud de los pacientes y mejoró la personalización de los tratamientos, añadiendo un valor significativo al estudio. No se identificaron impactos negativos no previstos, ya que el diseño metodológico riguroso y la constante evaluación de los impactos potenciales

aseguraron que cualquier efecto adverso potencial fuese mitigado adecuadamente. En resumen, el proyecto no solo ha alcanzado sus objetivos previstos en términos de sostenibilidad, ética y diversidad, sino que también ha generado impactos positivos adicionales no previstos que han enriquecido los hallazgos y su aplicabilidad en el ámbito clínico.

6. Glosario

Análisis Exploratorio de Datos (EDA): Técnica utilizada para analizar conjuntos de datos y resumir sus características principales, a menudo empleando métodos visuales.

CD4A_0: Variable objetivo en este estudio, que representa el recuento inicial de células CD4 en pacientes con VIH, utilizado como indicador del estado inmunológico.

Dimensionalidad: Número de variables o características en un conjunto de datos. Reducir la dimensionalidad implica disminuir el número de variables para simplificar el análisis sin perder información significativa.

Imputación Múltiple: Técnica para manejar datos faltantes que crea múltiples versiones imputadas del conjunto de datos y combina los resultados de cada uno para obtener estimaciones precisas y menos sesgadas.

PCA (Principal Component Analysis): Técnica estadística de reducción de dimensionalidad que transforma variables correlacionadas en un conjunto de valores de variables no correlacionadas llamadas componentes principales.

Random Forest: Algoritmo de aprendizaje automático utilizado para clasificación y regresión que construye múltiples árboles de decisión durante el entrenamiento y genera la media de las predicciones de los árboles individuales.

RMSE (Root Mean Square Error): Medida de la diferencia entre los valores predichos por un modelo y los valores observados, utilizada para evaluar la precisión de un modelo predictivo.

VIH (Virus de la Inmunodeficiencia Humana): Virus que causa el síndrome de inmunodeficiencia adquirida (SIDA), afectando el sistema inmunológico y reduciendo la capacidad del cuerpo para combatir infecciones y enfermedades.

ODS (Objetivos de Desarrollo Sostenible): Conjunto de 17 objetivos globales adoptados por las Naciones Unidas para abordar desafíos globales, incluyendo la pobreza, la desigualdad, el cambio climático, la degradación ambiental, la paz y la justicia.

Técnicas de Imputación: Métodos para manejar datos faltantes en conjuntos de datos, incluyendo imputación por la media, mediana, moda, imputación múltiple y métodos avanzados como la imputación iterativa.

7. Bibliografía

- [1] Fleming, T. R. (2011). Addressing missing data in clinical trials. *Annals of Internal Medicine*.
- [2] Haukoos, J. S., & Newgard, C. D. (2007). Advanced statistics: Missing data in clinical research - Part 1: An introduction and conceptual framework. *Academic Emergency Medicine*.
- [3] Kenward, M. G. (2013). The handling of missing data in clinical trials. *Clinical Investigation*.
- [4] Stack, C. B., Butterworth, T., & Goldin, R. (2018). Designed Learning: Missing Data in Clinical Research. *Annals of Internal Medicine*.
- [5] Kagan, J. M., Sanchez, A. M., Landay, A. L., & Denny, T. N. (2015). A Brief Chronicle of CD4 as a Biomarker for HIV/AIDS: A Tribute to the Memory of John L. Fahey. *Forum on Immunopathological Diseases and Therapeutics*.
- [6] Bello, V. (2023). Expanding HIV Clinical Monitoring: The Role of CD4, CD8, and CD4/CD8 Ratio in Predicting Non-AIDS Events.
- [7] Lembas, A., Załęski, A., Mikula, T., Dyda, T., Stańczak, W., & Wiercińska-Drapała, A. (2022). Evaluation of Clinical Biomarkers Related to CD4 Recovery in HIV-Infected Patients—5-Year Observation. *Viruses*.
- [8] Zhao, X., Han, B., Zhong, J., Lindborg, S., Thomas, N., & Liu, G. F. (2019). Handling Missing Data in Clinical Trials with Bayesian and Frequentist Approaches.
- [9] Little, R. J., & Rubin, D. B. (2019). *Statistical analysis with missing data*. John Wiley & Sons.
- [10] Graham, J. W. (2009). Missing data: methods and practice. *Annual Review of Psychology*.
- [11] Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons.
- [12] Van Buuren, S. (2018). *Flexible Imputation of Missing Data*. CRC Press.
- [13] Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D., & Altman, R. B. (2001). Missing value estimation methods for DNA microarrays. *Bioinformatics*.
- [14] Sterne, J. A., White, I. R., Carlin, J. B., Spratt, M., Royston, P., Kenward, M. G., Wood, A. M., & Carpenter, J. R. (2009). Addressing missing data in randomized trials: a practical guide. *BMJ*.
- [15] Enders, C. K. (2010). *Applied Missing Data Analysis*. Guilford Press.
- [16] Echeverría, P., Negredo, E., Carosi, G., Gálvez, J., Gómez, J. L., Ocampo, A., Portilla, J., Prieto, A., López, J. C., Rubio, R., Mariño, A., Pedrol, E., Viladés,

- C., del Arco, A., Moreno, A., Bravo, I., López-Blazquez, R., Pérez-Alvarez, N., & Clotet, B. (2010). Similar antiviral efficacy and tolerability between efavirenz and lopinavir/ritonavir, administered with abacavir/lamivudine (Kivexa), in antiretroviral-naïve patients: a 48-week, multicentre, randomized study (Lake Study). *Antiviral research*, 85(2), 403–408. <https://doi.org/10.1016/j.antiviral.2009.11.008>
- [17] Little, R. J., & Rubin, D. B. (2019). Statistical analysis with missing data. John Wiley & Sons.
- [18] Graham, J. W. (2009). Missing data: methods and practice. *Annual Review of Psychology*.
- [19] Van Buuren, S. (2018). Flexible Imputation of Missing Data. CRC Press.
- [20] Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D., & Altman, R. B. (2001). Missing value estimation methods for DNA microarrays. *Bioinformatics*.
- [21] Zhao, Z., Liu, R., Groner, J. I., Xiang, H., & Zhang, P. (2023). Data Imputation for Clinical Trial Emulation: A Case Study on Impact of Intracranial Pressure Monitoring for Traumatic Brain Injury. *medRxiv*.
- [22] Sterne, J. A., White, I. R., Carlin, J. B., Spratt, M., Royston, P., Kenward, M. G., Wood, A. M., & Carpenter, J. R. (2009). Addressing missing data in randomized trials: a practical guide. *BMJ*.
- [23] Gurpreet Kaur y Rani, R. (2022). An Efficient Predictive Model for High Dimensional Data. Springer.
- [24] Jotheeswaran, J., & Koteeswaran, S. (2016). Feature Selection using Random Forest Method for Sentiment Analysis. *Indian Journal of Science and Technology*.
- [25] Gharsalli, S., Emile, B., Laurent, H., & Desquesnes, X. (2016). Feature Selection for Emotion Recognition based on Random Forest. *International Conference on Computer Vision Theory and Applications*.
- [26] Wang, Y., & Xia, S.-T. (2016). A novel feature subspace selection method in random forests for high dimensional data. *International Joint Conference on Neural Network*.
- [27] Liu, A., Liu, C., Deng, X., Huang, Y., Liao, L., Meng, Z., He, M., & Huang, J. (2021). The association between serum CD4 T lymphocyte counts and surgical outcomes in HIV/AIDS patients in Guangxi, China: a retrospective cohort study. *PeerJ*.
- [28] Wu, Y. (2023). Effects of Gender and Baseline CD4 Count on Post-Treatment CD4 Count Recovery and Outcomes in Patients with Advanced HIV Disease: A Retrospective Cohort Study. *AIDS Research and Human Retroviruses*.
- [29] Grover, S., Mehta, P., Wang, Q., Bhatia, R., Bvochora-Nsingo, M., Davey, S., Iyengar, M. F., Shah, S., Shin, S. S., & Zetola, N. M. (2020). Association Between CD4 Count and Chemoradiation Therapy Outcomes Among Cervical Cancer Patients With HIV. *Journal of Acquired Immune Deficiency Syndromes*.

- [30] Patel, K. C., Patel, V., & Samnani, S. (2022). Evaluation of CD4 count response in HIV subjects with antiretroviral treatment protocol. *Indian Journal of Immunology and Respiratory Medicine*.
- [31] Khat, N., Kudligi, C., Rathod, R. M., Kuntoji, V., Bhagwat, P. V., Ramachandra, S. T., & Chavan, R. L. (2020). A clinical study of mucocutaneous manifestation of HIV/AIDs and its correlation with CD4 count. *Journal of Pakistan Association of Dermatology*. [32] Zhang, L., Cui, H., & Welsch, R. E. (2020, October). A Study on Multidimensional Medical Data Processing Based on Random Forest. In *2020 5th International Conference on Universal Village (UV)* (pp. 1-5). IEEE.
- [33] Tozan Rüzgar, Ş. (2022). Data Exploration. In *Practical Data Science with Python*
- [34] Ma, Y., & Zhu, L. (2013). A review on dimension reduction. *International Statistical Review*, 81(1), 134-150.
- [35] Dhal, S., Sauna, T., & Aune, Y. (2018). Lipid abnormalities and HIV disease progression: A systematic review and meta-analysis. *Journal of Acquired Immune Deficiency Syndromes*, 77(1), 1-11.
- [36] Bariha, D., Jandu, R., & John, T. (2018). CD4 count and its clinical implications in HIV infection: A comprehensive review. *Journal of Clinical and Cellular Immunology*, 108(1), 1-8.
- [37] Dell'Aira, F. (2022). Exploratory data analyses: describing our data. In *Essentials of Bioinformatics*
- [38] Aliyannissa, A., Kuswiyanto, R. B., Setiabudi, D., Nataprawira, H. M., Alam, A., & Sekarwana, N. (2020). Correlation between CD4 count and glomerular filtration rate or urine protein
- [39] Dhal, N., Panda, S., Mohapatra, N., Pattanayak, N. C., & Pattanaik, R. (2018). Study of haematological abnormalities in HIV infected patients and its correlation with CD4 counts. *International Journal of Research in Medical Sciences*.
- [40] Bariha, P. K., Karua, P. C., & Tudu, K. M. (2018). Correlations between clinical features and CD4 cell count in HIV patients with tuberculosis. *International Journal of Advances in Medicine*.
- [41] Khat, N., Kudligi, C., Rathod, R. M., Kuntoji, V., Bhagwat, P. V., Ramachandra, S. T., & Chavan, R. L. (2020). A clinical study of mucocutaneous manifestation of HIV/AIDs and its correlation with CD4 count. *Journal of Pakistan Association of Dermatology*.
- [42] Shafizadeh-Moghadam, H. (2021). Fully component selection: An efficient combination of feature selection and principal component analysis to increase model performance. *Expert Systems with Applications*, 186, 115678.
- [43] Tarchoune, I., Djebbar, A., Merouani, H. F., & Hadji, D. (2022). An improved random forest based on feature selection and feature weighting for case retrieval in CBR systems: application to medical data. *International Journal of Software Innovation (IJSI)*, 10(1), 1-20.

- [44] Zhang, L., Cui, H., & Welsch, R. E. (2020, October). A Study on Multidimensional Medical Data Processing Based on Random Forest. In 2020 5th International Conference on Universal Village (UV) (pp. 1-5). IEEE.
- [45] Mohapatra, S. K., & Mohanty, M. N. (2020). Big data analysis and classification of biomedical signal using random forest algorithm. In New paradigm in decision science and management: Proceedings of ICDSM 2018 (pp. 217-224). Springer Singapore.

8. Anexos

ANEXO 1

```

1 #Se verifica la cantidad de valores faltantes en la variable objetivo 'CD4A_0'
2 missing_values_target = data['CD4A_0'].isna().sum()
3
4 # Identificar variables categóricas en el dataset
5 categorical_vars = data.select_dtypes(include=['object']).columns.tolist()
6
7 missing_values_target, categorical_vars

(16,
 ['nusuario',
  'npac',
  'fecha_nac',
  'especificar',
  'fecha_ini_lake',
  'fecha_vih',
  'Fecha_0',
  'Fecha_12',
  'Fecha_24',
  'Fecha_36',
  'Fecha_48'])


```

ANEXO 2

```

1 # Calcular el porcentaje de valores NaN para cada variable numérica
2 nan_percentages_num = numeric_data.isna().mean() * 100
3
4 # Filtrar variables con más del 50% de valores NaN
5 high_nan_columns_num = nan_percentages_num[nan_percentages_num > 50].sort_values(ascending=False)
6
7 high_nan_columns_num

```

✓ 0.0s  Open 'high_nan_columns_num' in Data Wrangler

a19	100.000000
Embarazo_12	100.000000
Estado	100.000000
a28	100.000000
VHB_24	100.000000
...	
CD4P_36	50.862069
CD8A_36	50.862069
CD8P_36	50.862069
Urea_mg_36	50.862069
Trigliceridos_mg_36	50.862069

Length: 99, dtype: float64

ANEXO 3

```
1 # Estadísticas descriptivas
2 cd4a_0_descriptive_stats = numeric_data_cleaned['CD4A_0'].describe()
3
4 #distribución
5 fig, ax = plt.subplots(1, 2, figsize=(12, 5))
6
7 # Histograma
8 sns.histplot(numeric_data_cleaned['CD4A_0'].dropna(), bins=30, kde=True, ax=ax[0])
9 ax[0].set_title('Histograma de CD4A_0')
10 ax[0].set_xlabel('CD4A_0 (Absolutos)')
11 ax[0].set_ylabel('Frecuencia')
12
13 # Boxplot
14 sns.boxplot(x=numeric_data_cleaned['CD4A_0'].dropna(), ax=ax[1])
15 ax[1].set_title('Boxplot de CD4A_0')
16 ax[1].set_xlabel('CD4A_0 (Absolutos)')
17
18 plt.tight_layout()
19 plt.show()
20
21 # Mostrar estadísticas descriptivas
22 cd4a_0_descriptive_stats, numeric_data_cleaned['CD4A_0'].isna().sum()
```

ANEXO 4

```
1 #matriz de correlación centrada en la variable objetivo
2 correlation_matrix = numeric_data_cleaned.corr()['CD4A_0'].sort_values(ascending=False)
3
4 # ignoramos valores muy cercanos a cero
5 significant_correlations = correlation_matrix[abs(correlation_matrix) > 0.2]
6
7 plt.figure(figsize=(10, 8))
8 sns.heatmap(significant_correlations.to_frame(), annot=True, cmap='coolwarm', fmt=".2f")
9 plt.title('Mapa de calor de las correlaciones de CD4A_0 con otras variables')
10 plt.show()
11
12 significant_correlations
```

```
CD4A_0      1.000000
CD4A_24     0.690517
CD4P_12     0.617903
CD4P_24     0.605245
CD4P_0      0.510342
LinfosTotales_24 0.397907
Hemoglobina_0 0.372762
Urea_mg_0   0.299065
Hematocrito_0 0.298836
pH_12       0.290907
CD8A_0      0.253599
Plaquetas_24 0.248675
LinfosTotales_12 0.227594
CD4A_12     0.204924
Trigliceridos_mg_0 -0.211624
edad        -0.223126
Trigliceridos_mg_12 -0.242627
Colesterol_mg_12 -0.263242
CD8P_12     -0.270162
CD8P_24     -0.278808
estadio_VIH_31 -0.786699
Name: CD4A_0, dtype: float64
```


ANEXO 5

```
1 # Imputación de valores faltantes con la mediana, excluyendo la variable objetivo
2 X = numeric_data_cleaned.drop(columns=['CD4A_0', 'nvisita', 'edad'])
3 imputer = SimpleImputer(strategy='median')
4 X_imputed = pd.DataFrame(imputer.fit_transform(X), columns=X.columns)
5
6 # Normalización de los datos utilizando StandardScaler, excluyendo 'nvisita' y 'edad'
7 scaler = StandardScaler()
8 X_scaled_excl_nvisita_edad = pd.DataFrame(scaler.fit_transform(X_imputed), columns=X_imputed.columns)
9
10 # Reintegrar 'nvisita', 'edad' y la variable objetivo 'CD4A_0'
11 X_scaled_excl_nvisita_edad['nvisita'] = numeric_data_cleaned['nvisita'].values
12 X_scaled_excl_nvisita_edad['edad'] = numeric_data_cleaned['edad'].values
13 data_normalized_with_target_adjusted = X_scaled_excl_nvisita_edad.copy()
14 data_normalized_with_target_adjusted['CD4A_0'] = numeric_data_cleaned['CD4A_0']
15
16 # Comprobar la combinación y la integridad de los datos
17 na_count_target_after = data_normalized_with_target_adjusted['CD4A_0'].isna().sum()
18 data_normalized_with_target_adjusted.head(), na_count_target_after
```

ANEXO 6

```
1 plt.figure(figsize=(15, 10))
2 for i, var in enumerate(X_scaled_excl_nvisita_edad.columns[:10], 1):
3     plt.subplot(2, 5, i)
4     sns.histplot(X_scaled_excl_nvisita_edad[var], kde=True)
5     plt.title(f'Histograma de {var} (Normalizado)')
6 plt.tight_layout()
7 plt.show()

✓ 1.4s
```

ANEXO 7

```
1 # Se excluye la variable objetivo antes de aplicar PCA
2 X_pca_input = data_normalized_with_target_adjusted.drop(columns=['CD4A_0'])
3 # Asegurarnos de imputar valores faltantes en las características, excluyendo la variable objetivo
4 imputer = SimpleImputer(strategy='median')
5 X_pca_input_imputed = pd.DataFrame(imputer.fit_transform(X_pca_input), columns=X_pca_input.columns)
6
7 pca = PCA()
8 X_pca = pca.fit_transform(X_pca_input_imputed)
9
10 # Varianza explicada por cada componente
11 explained_variance = pca.explained_variance_ratio_
12
13 # varianza acumulada
14 cumulative_explained_variance = explained_variance.cumsum()
15
16 # Determinar el número de componentes necesarios para explicar al menos el 95% de la varianza
17 num_components_95 = next(i for i, total_var in enumerate(cumulative_explained_variance) if total_var >= 0.95) + 1
18
19 num_components_95

✓ 0.0s
```

40

```
1 # Aplicar PCA con el número de componentes determinado
2 pca = PCA(n_components=num_components_95)
3 X_pca_transformed = pca.fit_transform(X_pca_input_imputed)
4
5 # Crear un DataFrame con los componentes principales
6 X_pca_df = pd.DataFrame(X_pca_transformed, columns=[f'PC{i+1}' for i in range(num_components_95)])
7
8 # Añadir la variable objetivo al DataFrame de componentes principales
9 X_pca_df['CD4A_0'] = data_normalized_with_target_adjusted['CD4A_0'].values
10
11 # Verificar la transformación
12 X_pca_df.head()
```

ANEXO 8

```
1 df_dropped = X_pca_df.dropna(subset=['CD4A_0'])
2
3 # Separar las características y la variable objetivo
4 X_dropped = df_dropped.drop(columns=['CD4A_0'])
5 y_dropped = df_dropped['CD4A_0']
6
7 # Entrenar y evaluar el modelo de Random Forest
8 rf_model_dropped = RandomForestRegressor(n_estimators=100, random_state=42)
9 scores_dropped = cross_val_score(rf_model_dropped, X_dropped, y_dropped, cv=5, scoring='neg_mean_squared_error')
10 rmse_scores_dropped = (-scores_dropped) ** 0.5
11
12 # Resultados
13 print(f'Eliminación de Filas - RMSE medio: {rmse_scores_dropped.mean():.2f}, Desviación estándar: {rmse_scores_dropped.std():.2f}')
✓ 1.0s
Eliminación de Filas - RMSE medio: 101.94, Desviación estándar: 15.44

1 # Imputar los valores faltantes en la variable objetivo con la mediana
2 imputer_median = SimpleImputer(strategy='median')
3 X_pca_df['CD4A_0_imputed'] = imputer_median.fit_transform(X_pca_df[['CD4A_0']])
4
5 # Separar las características y la variable objetivo
6 X_imputed = X_pca_df.drop(columns=['CD4A_0'])
7 y_imputed = X_pca_df['CD4A_0_imputed']
8
9 # Random Forest
10 rf_model_imputed = RandomForestRegressor(n_estimators=100, random_state=42)
11 scores_imputed = cross_val_score(rf_model_imputed, X_imputed, y_imputed, cv=5, scoring='neg_mean_squared_error')
12 rmse_scores_imputed = (-scores_imputed) ** 0.5
13
14 # Resultados
15 print(f'Imputación Simple (Mediana) - RMSE medio: {rmse_scores_imputed.mean():.2f}, Desviación estándar: {rmse_scores_imputed.std():.2f}')
✓ 1.1s
Imputación Simple (Mediana) - RMSE medio: 14.69, Desviación estándar: 9.04

1 # Dividir el conjunto de datos en casos completos y casos faltantes
2 complete_cases = X_pca_df.dropna(subset=['CD4A_0'])
3 missing_cases = X_pca_df[X_pca_df['CD4A_0'].isna()]
4
5 # modelo para predecir los valores faltantes
6 X_complete = complete_cases.drop(columns=['CD4A_0'])
7 y_complete = complete_cases['CD4A_0']
8 rf_prelim_model = RandomForestRegressor(n_estimators=100, random_state=42)
9 rf_prelim_model.fit(X_complete, y_complete)
10
11 # Predecir los valores faltantes
12 X_missing = missing_cases.drop(columns=['CD4A_0'])
13 y_missing_pred = rf_prelim_model.predict(X_missing)
14
15 # Imputar los valores predichos
16 X_pca_df.loc[X_pca_df['CD4A_0'].isna(), 'CD4A_0'] = y_missing_pred
17
18 # Separar las características y la variable objetivo
19 X_modeled = X_pca_df.drop(columns=['CD4A_0'])
20 y_modeled = X_pca_df['CD4A_0']
21
22 # Entrenar y evaluar el modelo de Random Forest
23 rf_model_modeled = RandomForestRegressor(n_estimators=100, random_state=42)
24 scores_modeled = cross_val_score(rf_model_modeled, X_modeled, y_modeled, cv=5, scoring='neg_mean_squared_error')
25 rmse_scores_modeled = (-scores_modeled) ** 0.5
26
27 # Resultados
28 print(f'Modelado para Imputación - RMSE medio: {rmse_scores_modeled.mean():.2f}, Desviación estándar: {rmse_scores_modeled.std():.2f}')
✓ 13s
Modelado para Imputación - RMSE medio: 15.03, Desviación estándar: 9.90

1 # Imputación múltiple
2 iterative_imputer = IterativeImputer(random_state=42)
3 X_pca_df['CD4A_0_imputed_multiple'] = iterative_imputer.fit_transform(X_pca_df[['CD4A_0']])
4
5 # Separar las características y la variable objetivo
6 X_imputed_multiple = X_pca_df.drop(columns=['CD4A_0'])
7 y_imputed_multiple = X_pca_df['CD4A_0_imputed_multiple']
8
9 # Entrenar y evaluar el modelo de Random Forest
10 rf_model_imputed_multiple = RandomForestRegressor(n_estimators=100, random_state=42)
11 scores_imputed_multiple = cross_val_score(rf_model_imputed_multiple, X_imputed_multiple, y_imputed_multiple, cv=5, scoring='neg_mean_squared_error')
12 rmse_scores_imputed_multiple = (-scores_imputed_multiple) ** 0.5
13
14 # Resultados
15 print(f'Imputación Múltiple - RMSE medio: {rmse_scores_imputed_multiple.mean():.2f}, Desviación estándar: {rmse_scores_imputed_multiple.std():.2f}')
✓ 1.1s
Imputación Múltiple - RMSE medio: 13.20, Desviación estándar: 8.50
```

```

1 # Datos para el gráfico
2 methods = ['Eliminación de Filas', 'Imputación Mediana', 'Modelado para Imputación', 'Imputación Múltiple']
3 rmse_means = [rmse_scores_dropped.mean(), rmse_scores_imputed.mean(), rmse_scores_modeled.mean(), rmse_scores_imputed_multiple.mean()]
4 rmse_stds = [rmse_scores_dropped.std(), rmse_scores_imputed.std(), rmse_scores_modeled.std(), rmse_scores_imputed_multiple.std()]
5
6 # gráfico de barras
7 plt.figure(figsize=(10, 6))
8 bars = plt.bar(methods, rmse_means, yerr=rmse_stds, capsize=5, color='skyblue')
9
10 # etiquetas y título
11 plt.xlabel('Métodos de Tratamiento de Missings')
12 plt.ylabel('RMSE Medio')
13 plt.title('Comparación de Métodos de Tratamiento de Datos Faltantes')
14 plt.ylim(0, max(rmse_means) + max(rmse_stds) + 10)
15
16 for bar, mean, std in zip(bars, rmse_means, rmse_stds):
17     yval = bar.get_height()
18     plt.text(bar.get_x() + bar.get_width()/2.0, yval + 1, f'{mean:.2f} ± {std:.2f}', ha='center', va='bottom')
19
20 plt.show()
21
✓ 0.1s

```

ANEXO 9

Archivo TFM.ipynb con el código utilizado