Universidad Europea

UNIVERSIDAD EUROPEA DE MADRID

ESCUELA DE ARQUITECTURA, INGENIERÍA Y DISEÑO

MÁSTER UNIVERSITARIO EN ANÁLISIS DE DATOS MASIVOS

TRABAJO FIN DE MÁSTER

GENERACIÓN DE DATOS SINTÉTICOS REALISTAS PARA REGISTROS CLÍNICOS: UN ENFOQUE BASADO EN CLUSTERING

GERARD MARTURIÀ I NAVARRO

Dirigido por

RICARD GAVALDÀ y ÁLVARO CALLE CORDÓN

CURSO 2023 - 2024

TÍTULO: GENERACIÓN DE DATOS SINTÉTICOS REALISTAS PARA REGISTROS CLÍNICOS: UN ENFOQUE BASADO EN CLUSTERING

AUTOR: GERARD MARTURIÀ I NAVARRO

TITULACIÓN: MÁSTER UNIVERSITARIO EN ANÁLISIS DE DATOS MASIVOS

DIRECTOR DEL PROYECTO: RICARD GAVALDÀ y ÁLVARO CALLE CORDÓN

FECHA: Septiembre de 2024

RESUMEN

Este proyecto aborda el desarrollo de un sistema para la generación de datos sintéticos realistas a partir de registros clínicos reales, con el objetivo de garantizar la privacidad de los pacientes y, al mismo tiempo, preservar la utilidad y representatividad de los datos. En el ámbito de la salud, la disponibilidad de datos reales está sujeta a estrictas normativas de confidencialidad, lo que dificulta la realización de estudios estadísticos conjuntos entre instituciones. En este contexto, los datos sintéticos ofrecen una alternativa viable, permitiendo realizar análisis y pruebas sin comprometer la integridad de la información confidencial.

El proyecto se basó en la segmentación de episodios clínicos mediante técnicas de clustering aplicadas a un conjunto de datos anonimizados. Este proceso se llevó a cabo utilizando un enfoque basado en Naive Bayes Clustering, donde se identificaron patrones complejos en los datos y se agruparon episodios hospitalarios en clusters de pacientes que reflejan similitudes clínicas y demográficas. Posteriormente, se implementaron modelos generativos que replican las distribuciones de las variables para generar episodios sintéticos de alta calidad.

La validación de los datos sintéticos se llevó a cabo utilizando un modelo de Machine Learning basado en Random Forest, evaluando su capacidad para distinguir entre datos reales y sintéticos. Aunque el Random Forest logró distinguir entre ambos conjuntos de datos con cierto grado de acierto, el algoritmo desarrollado fue capaz de generar datos sintéticos altamente similares a los reales, representando así un avance significativo en este campo. Estos resultados sientan las bases para futuras investigaciones en la generación de datos sintéticos más sofisticados.

Los principales resultados de este proyecto incluyen el diseño y desarrollo de un algoritmo de generación de datos sintéticos y la implementación de un sistema de validación para garantizar la coherencia y realismo de los datos generados. Estos avances tienen aplicaciones directas en áreas como la investigación clínica, el desarrollo de modelos predictivos y la planificación de recursos sanitarios.

Palabras clave: Datos sintéticos, privacidad de datos, aprendizaje automático, clustering, registros clínicos, Naive Bayes, Random Forest.

ABSTRACT

This project focuses on the development of a system for generating realistic synthetic data from real clinical records, aiming to ensure patient privacy while preserving the utility and representativeness of the data. In the healthcare sector, access to real data is restricted due to strict confidentiality regulations, which limits the possibility of joint statistical studies between institutions. In this context, synthetic data provides a viable alternative, allowing analysis and testing without compromising the integrity of confidential information.

The project employed clustering techniques to segment anonymized clinical episodes from a real dataset. Using a Naive Bayes-based Clustering approach, the project identified complex patterns in the data and grouped hospital episodes into clusters that reflect clinical and demographic similarities. Subsequently, generative models were implemented to replicate the variable distributions and produce high-quality synthetic episodes.

The validation of the synthetic data was performed using a Random Forest machine learning model, assessing its ability to distinguish between real and synthetic data. Although the Random Forest achieved a certain degree of accuracy in distinguishing between the two datasets, the developed algorithm was able to generate synthetic data that closely resembles the real data, marking a significant step forward in this domain. These results lay the groundwork for future research in generating more sophisticated synthetic data.

The main outcomes of this project include the design and development of a synthetic data generation algorithm and the implementation of a validation system to ensure the coherence and realism of the generated data. These advancements have direct applications in areas such as clinical research, the development of predictive models, and healthcare resource planning.

Keywords: Synthetic data, data privacy, machine learning, clustering, clinical records, Naive Bayes, Random Forest.

Agradecimientos

Me gustaría expresar mi más profundo agradecimiento a todas las personas que han sido parte esencial en la realización de este trabajo fin de máster. Su apoyo, guía y asesoramiento han sido fundamentales para que este proyecto se haya llevado a cabo con éxito.

Quiero dar un agradecimiento especial a mi tutor externo, Ricard Gavaldà, por su tiempo, compromiso y dedicación, y por compartir su vasta experiencia a lo largo de este proceso. Su disposición y valiosas orientaciones me han permitido sacar adelante este trabajo.

También quisiera expresar mi gratitud a mi tutor interno de la universidad, Álvaro Calle Cordón, por su constante revisión y su apoyo incondicional para asegurar que cumpliera con los plazos establecidos.

Finalmente, quiero agradecer a todas las personas que, de una u otra forma, han contribuido a la culminación de este proyecto. Su ayuda y apoyo han sido esenciales, y sin ustedes, esto no habría sido posible.

Gracias a todos.

Índice general

Índice de Figuras

Índice de Tablas

Capítulo 1. Introducción

Este capítulo presenta una visión general del trabajo realizado, siguiendo la estructura de las secciones incluidas.

1.1 Estructura de la memoria

A continuación se describe brevemente la estructura de la memoria del trabajo fin de máster, especificando el contenido de cada capítulo:

- **Capítulo 1: Introducción** - Se presenta una visión general del proyecto, incluyendo el contexto, los objetivos y la justificación del trabajo, además de un resumen de los principales resultados obtenidos.
- **Capítulo 2: Estado del arte** - Se realiza una revisión del estado del arte, describiendo los métodos y avances más relevantes en la generación de datos sintéticos y su aplicación en el ámbito de la salud, así como los desafíos existentes.
- **Capítulo 3: Objetivos** - En este capítulo se detallan los objetivos generales y específicos del proyecto, ajustados según el desarrollo del mismo, y los beneficios esperados para las instituciones sanitarias y la investigación médica.
- **Capítulo 4: Desarrollo del proyecto** - Se describe la planificación y las actividades llevadas a cabo durante el desarrollo del proyecto, desde la implementación del algoritmo hasta la evaluación de los resultados.
- **Capítulo 5: Metodología** - Se explica la metodología empleada para la generación de datos sintéticos, incluyendo el análisis de las variables, el algoritmo implementado y la validación de los datos generados.
- **Capítulo 6: Resultados** - Se presentan los resultados obtenidos en el proyecto, detallando la calidad de los datos sintéticos generados y los análisis realizados para validarlos.
- **Capítulo 7: Conclusiones** - Este capítulo resume las principales conclusiones del proyecto, resaltando los logros alcanzados.
- **Capítulo 8: Discusión y trabajo futuro** - Se discuten las limitaciones encontradas, posibles mejoras y líneas de trabajo futuro para seguir desarrollando y perfeccionando los métodos empleados en este proyecto.

1.2 Contexto y justificación

En el ámbito del Big Data, la generación de datos sintéticos se ha vuelto esencial debido a la necesidad de disponer de conjuntos de datos realistas y representativos para probar y evaluar algoritmos, sistemas y aplicaciones. Esto es especialmente crítico en áreas como la salud, donde la confidencialidad y sensibilidad de los datos reales imponen restricciones significativas. La generación de datos sintéticos permite realizar investigaciones y análisis sin comprometer la privacidad de los pacientes, lo cual es crucial en entornos donde las restricciones legales o éticas limitan el uso compartido de datos reales. Además, los datos sintéticos son útiles para probar nuevos algoritmos o sistemas en escenarios hipotéticos, permitiendo una preparación proactiva para futuras necesidades de tratamiento.

1.3 Planteamiento del problema

A pesar de la utilidad de los datos sintéticos, crear conjuntos de datos que sean precisos y reflejen fielmente la diversidad y las características estadísticas de las poblaciones reales sigue siendo un desafío. Estos problemas son críticos, ya que los datos inexactos pueden conducir a conclusiones erróneas en estudios clínicos y análisis de políticas de salud.

Mi supervisor, R. Gavaldà, me propuso seguir la metodología del trabajo de Aviñó *et al.* [1], donde se trabajaba con pacientes y episodios de manera conjunta, para investigar si la clusterización de pacientes permitiría generar episodios realistas. En este proyecto, la clusterización se realizó a nivel de pacientes, pero la generación de datos tenía que ser a nivel de episodios. Esta aproximación buscaba determinar si, al agrupar pacientes con características similares, podríamos generar episodios que reflejaran fielmente la realidad clínica.

Inicialmente, se contempló la federación como un medio para adaptar estos procesos a múltiples fuentes hospitalarias. Sin embargo, en este proyecto se ha optado por un enfoque más flexible que permite la aplicación del algoritmo en un conjunto de datos homogéneo, garantizando la adaptabilidad y utilidad en distintos entornos sin la necesidad de implementar una federación compleja. Esto facilita la generación de datos sintéticos en contextos donde los datos tienen un formato común en los sistemas de información hospitalarios.

1.4 Objetivos del proyecto

El objetivo general de este trabajo es desarrollar y validar una metodología para la generación de datos sintéticos en el ámbito de la salud, enfocándose en episodios clínicos de pacientes agudos. Esta metodología se basa en el análisis exhaustivo de un conjunto de datos real, utilizando técnicas de clustering con modelos Naive Bayes, basadas en los trabajos previos de Ruffini *et al.* [2, 3], para segmentar los datos originales y posteriormente generar conjuntos de datos sintéticos. La similitud entre los datos sintéticos y los datos reales se evalúa mediante métricas estadísticas y modelos de clasificación, como Random Forest, con el fin de determinar la capacidad de distinguir entre ambos conjuntos de datos, siguiendo la metodología propuesta en el trabajo previo de Aviñó *et al.* [1].

Los objetivos específicos son:

- Generar datos sintéticos que reflejen las características de los datos originales para facilitar su uso sin comprometer la privacidad.
- Validar objetivamente la calidad de los datos sintéticos mediante un modelo de Random Forest [4] para distinguir entre datos reales y sintéticos.
- Asegurar la representatividad y precisión de los datos sintéticos para reflejar la variabilidad observada en los datos originales.
- Cumplir con las normativas de privacidad al reducir la necesidad de compartir datos sensibles mediante el uso de datos sintéticos.
- Facilitar investigaciones futuras permitiendo la prueba de nuevas tecnologías y sistemas con datos sintéticos representativos en entornos controlados.

La consecución de estos objetivos permitirá a las instituciones sanitarias disponer de una herramienta para generar datos sintéticos de calidad, que puedan utilizarse en diversos análisis

y desarrollos, fomentando la investigación y colaboración sin vulnerar la confidencialidad de los datos personales.

1.5 Metodología General

El proceso se inicia con un análisis exploratorio de las variables presentes en el conjunto de datos real, seguido por la preparación de los datos para su posterior agrupamiento. Para este fin, se implementa un algoritmo de clustering basado en Naive Bayes, con el objetivo de segmentar los episodios hospitalarios en *clusters* homogéneos. Posteriormente, se calculan métricas específicas para cada *cluster*, que se utilizan para generar datos sintéticos que replican las características observadas en los datos originales. Finalmente, la calidad y similitud de los datos sintéticos con respecto a los datos reales se evalúan mediante un modelo de Random Forest.

1.6 Resultados Obtenidos

Los resultados de este estudio incluyen la implementación efectiva de un algoritmo para la generación de datos sintéticos a partir de un conjunto de datos homogéneos de salud. Los datos generados presentan características estadísticas similares a las de los datos reales, lo cual ha sido validado mediante un conjunto de métricas y análisis detallados. Sin embargo, se observó que el modelo de Random Forest es capaz de diferenciar con relativa precisión entre datos reales y sintéticos. Esta capacidad de detección puede atribuirse a la dificultad del algoritmo para capturar la complejidad completa de las relaciones y la secuencia de eventos presentes en los episodios clínicos.

A pesar de esta limitación, los resultados obtenidos demuestran la eficacia de la metodología desarrollada y su potencial aplicabilidad en contextos sanitarios, sirviendo como base para futuras mejoras en la generación de datos sintéticos que puedan replicar de manera más efectiva la complejidad de los sucesos clínicos.

Capítulo 2. Estado del Arte

Este capítulo contextualiza el proyecto y presenta los aspectos necesarios para comprender las alternativas estudiadas.

2.1 Estado del Arte

En el ámbito de la salud, la generación y análisis de datos sintéticos a partir de registros médicos reales presentan desafíos significativos, principalmente relacionados con la privacidad de los pacientes y la representatividad de los datos [5]. Estos desafíos se amplifican cuando se considera la federación de datos provenientes de múltiples fuentes hospitalarias, cada una con sus propias características y distribuciones [6]. La necesidad de conjuntos de datos sintéticos que imiten fielmente los casos reales es evidente en numerosos campos de la ciencia y la ingeniería, ya que son cruciales para la evaluación de algoritmos y sistemas de software y hardware [7, 8]. Los datos sintéticos permiten establecer puntos de referencia comparativos y facilitan la extrapolación de resultados a futuros escenarios donde los datos reales no están disponibles o son insuficientes. Además, los generadores de datos sintéticos, al ajustar los parámetros de síntesis, permiten la exploración controlada de situaciones para las cuales no se dispone de datos reales, facilitando la investigación de relaciones causales en casos donde la intervención directa es inviable.

Tradicionalmente, en el ámbito sanitario, se han utilizado métodos como simuladores basados en el conocimiento humano para crear datos sintéticos. Sin embargo, con el avance de las técnicas de aprendizaje automático (ML, por sus siglas en inglés), estos procesos se han automatizado mediante enfoques como las Redes Generativas Adversarias (GANs, por sus siglas en inglés). A pesar de su éxito en la generación de imágenes realistas, como se aprecia en el estudio [9], las GANs presentan problemas como el colapso de modos y la falta de interpretabilidad, lo que las hace menos adecuadas en contextos sanitarios [10].

La generación de datos tabulares, como los registros clínicos, sigue presentando desafíos específicos. Las GANs han sido aplicadas para generar series temporales de datos médicos, simular los efectos de tratamientos individualizados y reducir el sesgo en los conjuntos de datos de entrenamiento [11]. Sin embargo, la naturaleza mixta de datos continuos y categóricos en los registros médicos dificulta su modelado efectivo.

En años recientes, el uso de GANs ha ganado tracción en la generación de datos tabulares. Uno de los avances más importantes en esta área ha sido el desarrollo de Conditional Tabular GAN (CTGAN), como se describe en [12], que aborda las dificultades de modelado en columnas continuas multimodales y columnas categóricas desbalanceadas. En pruebas realizadas, CTGAN superó a las redes bayesianas en la mayoría de los conjuntos de datos reales, consolidándose como un modelo clave en la generación de datos tabulares sintéticos.

Además, el estudio [13] revisó el uso de GANs para la síntesis de datos tabulares en el contexto de sistemas de detección de intrusiones (IDS). Este trabajo sugiere que la selección adecuada del modelo es fundamental para capturar patrones complejos y variados en los datos, lo que tiene implicaciones no solo en la ciberseguridad, sino también en áreas como la salud.

No obstante, las GANs enfrentan dificultades como la incapacidad de capturar adecuadamente la diversidad de los datos originales, lo cual es crucial en aplicaciones sanitarias. Este fenómeno, conocido como *colapso de modos*, ocurre cuando las GANs generan solo algunas de las posibles variaciones dentro de la distribución subyacente, como se señala en los estudios [9, 14]. Esto puede ser problemático en contextos sanitarios, donde la diversidad de los datos es esencial para obtener conclusiones válidas [15].

Además, los enfoques basados en GANs presentan problemas de escasa interpretabilidad, ya que se basan en redes neuronales profundas. Esto dificulta la evaluación de las razones detrás de la generación de ciertas muestras y la interpretación clara del modelo generativo subyacente. Este aspecto es especialmente crítico en el ámbito sanitario, donde es fundamental comprender y justificar las decisiones tomadas por los modelos, tanto para la confianza de los profesionales como para cumplir con regulaciones éticas y legales.

A pesar de los avances mencionados, la generación de datos sintéticos en el ámbito sanitario sigue enfrentando problemas de precisión y representatividad. Por ejemplo, el estudio [16] subraya la inexactitud de los códigos de la Clasificación Internacional de Enfermedades (CIE) en los resúmenes de alta hospitalaria como una medida de la ocurrencia de complicaciones en pacientes. Esto destaca la necesidad de metodologías más precisas para la generación y análisis de datos.

Como alternativa a las GANs, se han propuesto métodos basados en modelos generativos interpretables y técnicas estadísticas avanzadas. En particular, la descomposición de tensores y el método de momentos han demostrado ser útiles para modelar datos de salud de manera más precisa y comprensible. Investigaciones como las de [2, 3, 17] proponen la utilización de la descomposición de tensores para agrupar pacientes y analizar patrones complejos en los datos de salud, mejorando la precisión de los *clusters* generados y permitiendo una interpretación más clara de los resultados.

Estos métodos permiten capturar la estructura latente de los datos y generar conjuntos sintéticos que preservan las características estadísticas y la diversidad presente en los datos reales. Además, al ser modelos más interpretables, facilitan la comprensión y validación por parte de profesionales de la salud, lo cual es crucial para su aceptación y uso en entornos clínicos. Esto contrasta con los problemas de interpretabilidad asociados a las GANs, lo que ha generado un interés creciente en modelos generativos más transparentes.

Otras investigaciones, como la de [1], presentan un enfoque similar, pero centrado en la generación de registros médicos sintéticos mediante el uso de variables latentes y técnicas de descomposición de tensores. Este enfoque evitaba los problemas de colapso de modos observados en las GANs y proporcionaba una mayor interpretabilidad del modelo generativo. Los resultados de este estudio indicaron que los datos sintéticos eran plausibles y difíciles de distinguir de los reales mediante técnicas de clasificación como Random Forest. Además, este enfoque abordó de manera efectiva el problema de la privacidad de los pacientes, facilitando la colaboración entre instituciones hospitalarias sin necesidad de compartir datos reales.

2.2 Contexto y justificación

La necesidad de conjuntos de datos sintéticos que imiten fielmente los datos reales es crucial en el ámbito de la salud debido a la alta sensibilidad de los datos clínicos y las estrictas regulaciones de privacidad. Este proyecto busca abordar los desafíos relacionados con la generación y análisis de datos sintéticos, especialmente en la generación de episodios clínicos que reflejen adecuadamente las características y diversidad de los datos reales.

Al centrarse en episodios en lugar de pacientes individuales, se permite un análisis más detallado y específico de eventos clínicos, lo que es valioso para estudios epidemiológicos, evaluación de tratamientos y planificación de recursos sanitarios.

Este enfoque facilita la colaboración entre instituciones sanitarias, permitiendo compartir datos sintéticos generados localmente para realizar análisis conjuntos y comparativos sin revelar información sensible.

2.3 Planteamiento del problema

Aunque las Redes Generativas Adversariales (GANs) han mostrado ser eficaces para producir datos realistas en muchos contextos, presentan limitaciones como el colapso de modos y la falta de interpretabilidad, lo que restringe su uso en entornos médicos [9, 14, 18]. Estos problemas son especialmente relevantes en el análisis de episodios clínicos, donde la diversidad y complejidad de los eventos deben ser capturadas para garantizar resultados válidos y aplicables.

Alternativas como el uso de la Vault de Datos Sintéticos (SDV) han mostrado resultados prometedores en otros contextos, pero requieren ajustes significativos para cumplir con los requisitos y regulaciones específicas de los datos clínicos. Por lo tanto, surge la necesidad de desarrollar metodologías más robustas que aseguren la representatividad de los datos generados y permitan su interpretación adecuada.

En colaboración con la empresa Amalfi Analytics, este proyecto propone continuar la línea de investigación iniciada en estudios previos [1, 3], pero enfocándose en el análisis de episodios clínicos en lugar de pacientes individuales. Al adaptar técnicas como la descomposición tensorial y el uso de variables latentes para la generación de datos sintéticos de episodios clínicos, se espera superar algunas de las limitaciones observadas en las GANs, como el colapso de modos, y mejorar tanto la interpretación como la aplicabilidad de los datos generados.

Este enfoque permitirá no solo generar datos que preserven las características estadísticas de los episodios clínicos reales, sino también facilitar la colaboración entre instituciones hospitalarias. Los datos sintéticos así generados podrán utilizarse para estudios clínicos, planificación de recursos y análisis de tratamientos, sin comprometer la privacidad de los pacientes, proporcionando así un marco ético y eficaz para el uso de datos clínicos sintéticos en la investigación médica.

Capítulo 3. Objetivos

En este capítulo se incluye la descripción detallada de los objetivos del proyecto, basados en el análisis del estado del arte.

3.1 Objetivo general

El objetivo general de este trabajo es desarrollar y validar una metodología para la generación de datos sintéticos en el ámbito de la salud, enfocándose en episodios clínicos. Esta metodología se basa en el análisis exhaustivo de un conjunto de datos real, utilizando técnicas de clustering con modelos Naive Bayes para segmentar los datos originales, generar conjuntos de datos sintéticos que reproduzcan fielmente las características de los datos reales y evaluar la similitud mediante modelos de clasificación, como Random Forest, para determinar la capacidad de distinguir entre los datos reales y los sintéticos.

3.2 Objetivos específicos

1. **Análisis detallado del conjunto de datos real:**
    - Explorar y comprender profundamente el conjunto de datos de salud original, centrado en episodios clínicos, identificando variables clave y patrones que deben ser preservados en la generación de datos sintéticos.
2. **Aplicar técnicas de clustering basadas en Naive Bayes:**
    - Implementar un modelo de clustering basado en Naive Bayes para segmentar los episodios clínicos del conjunto de datos real, respetando las variaciones naturales y las distribuciones observadas.
3. **Generar datos sintéticos de episodios clínicos:**
    - Utilizar los parámetros y métricas obtenidos del análisis y clustering de los datos reales para generar conjuntos de datos sintéticos de episodios clínicos que reflejen las características y patrones observados.
4. **Evaluar la similitud entre los datos sintéticos y reales:**
    - Comparar las distribuciones estadísticas y frecuencias de los datos sintéticos y reales utilizando métricas adecuadas, analizando variables categóricas y continuas, para medir la similitud entre ambos conjuntos.
5. **Validar la calidad de los datos sintéticos mediante modelos de clasificación:**
    - Implementar un modelo de Random Forest para evaluar la capacidad de distinguir entre los datos sintéticos y reales. Una baja capacidad del modelo para diferenciar entre ambos indicará una alta calidad en los datos sintéticos generados.

3.3 Beneficios del Proyecto

Este proyecto tiene el potencial de ofrecer beneficios en el campo de la generación de datos sintéticos para el sector de la salud, contribuyendo con una metodología que puede facilitar el

análisis de datos sin comprometer la privacidad de los pacientes. Los beneficios específicos considerados incluyen:

- **Generación de datos sintéticos realistas:** La metodología desarrollada podría permitir la generación de datos sintéticos que reflejen de manera aproximada las características de los datos originales, lo que podría facilitar su uso en estudios y análisis sin la necesidad de compartir datos sensibles.
- **Validación objetiva de la calidad de los datos sintéticos:** El uso de un modelo de Random Forest para evaluar la capacidad de distinguir entre datos sintéticos y reales podría proporcionar una validación cuantitativa y objetiva de la calidad de los datos generados, aunque se reconoce que este método tiene sus limitaciones.
- **Representatividad y precisión:** Al comparar las frecuencias y distribuciones entre los datos reales y sintéticos, se espera que los conjuntos de datos sintéticos puedan reflejar de manera razonable la variabilidad observada en los originales, lo que podría mejorar la fiabilidad de los estudios futuros.
- **Cumplimiento de las normativas de privacidad:** El uso de datos sintéticos puede reducir la necesidad de compartir datos personales sensibles, contribuyendo al cumplimiento de normativas de privacidad como el GDPR, aunque es importante continuar evaluando la efectividad de estas medidas.
- **Facilitación de investigaciones futuras:** La capacidad de generar datos sintéticos representativos podría ser útil para que investigadores y desarrolladores prueben nuevas tecnologías, algoritmos y sistemas, fomentando la innovación en entornos seguros y controlados.

Capítulo 4. Desarrollo del proyecto

4.1 Planificación del Proyecto

La planificación del proyecto se dividió en varias fases principales, las cuales se describen a continuación:

- **Definición de objetivos y planificación (Abril):** Se llevaron a cabo reuniones iniciales con los tutores del proyecto para definir los objetivos y establecer el enfoque general. Además, se recibieron documentos clave, como el CMBD (Conjunto Mínimo Básico de Datos) anonimizado proporcionado por la empresa colaboradora Amalfi Analytics, para comenzar el estudio exploratorio de las variables. Paralelamente, se realizó un estudio exhaustivo del estado del arte, revisando literatura y trabajos previos que proporcionaron el marco teórico necesario para el desarrollo posterior.
- **Preparación y análisis de datos (Mayo-Junio):** Se trabajó en la preparación y análisis del conjunto de datos encriptado proporcionado por la empresa. Se desarrolló software específico para procesar los datos, aplicando algoritmos de clustering y cálculo de distribuciones. Esta fase fue esencial para establecer las bases del modelo generativo.
- **Desarrollo intensivo y evaluación (Julio-Agosto):** Se desarrollaron y ajustaron los algoritmos de generación de datos sintéticos. Durante este periodo, se refinaron los modelos comparando los datos sintéticos con los reales, asegurando que los resultados fueran precisos y útiles para cumplir los objetivos del proyecto.
- **Redacción y conclusiones (Septiembre):** Se redactó la memoria final del proyecto, integrando los resultados obtenidos. Se discutieron las limitaciones encontradas, se propusieron futuras líneas de trabajo y se presentaron las conclusiones basadas en los análisis realizados.

A lo largo del proyecto, se realizaron actividades complementarias como la visualización de material adicional para mejorar el contexto y reuniones periódicas con el equipo de Amalfi y asesores académicos. Estas reuniones permitieron ajustar el enfoque y avanzar de manera efectiva hacia los objetivos planteados.

Cronograma de Actividades del Proyecto

A continuación, se presenta un cronograma detallado de las actividades realizadas durante el proyecto, junto con el tiempo dedicado a cada una de ellas:

Tarea	Abril	Mayo	Junio	Julio	Agosto	Septiembre
Definición objetivos y enfoque	X					
Estudio del estado del arte	X	X				
Recepción y análisis de documentos clave (CMBD, estudios previos)	X					
Recepción y análisis del set de datos encriptado		X	X			
Desarrollo de software para procesamiento de datos		X	X			
Algoritmos de lectura, procesamiento, clusterización y creación de datos sintéticos		X	X			
Desarrollo de algoritmos de generación de datos sintéticos				X	X	
Ajuste y comparación de datos sintéticos con datos reales				X	X	
Redacción de la memoria final						X
Discusión de resultados y trabajo futuro						X
Reuniones periódicas con el equipo de Amalfi y asesores académicos	X	X	X	X	X	X

Tabla 1. Cronograma y descripción de actividades del proyecto.

Capítulo 5. Metodología

Este proyecto abarca una serie de pasos meticulosamente planificados que van desde el análisis inicial de las variables hasta la generación y validación de datos sintéticos. A continuación, se describe cada etapa del proceso, las metodologías empleadas, las herramientas utilizadas y los algoritmos desarrollados para cumplir con los objetivos planteados.

5.1 Análisis Inicial de Variables

El primer paso en el desarrollo del proyecto fue el análisis exhaustivo del conjunto de datos original, que contenía una gran cantidad de episodios clínicos registrados para cada paciente. Cada fila del conjunto de datos representaba un episodio o estancia hospitalaria de un paciente, y las columnas correspondían a diversas variables que describen aspectos clínicos y administrativos del tratamiento recibido.

El conjunto de datos se estructuró de acuerdo con las especificaciones del *Conjunto Mínimo Básico de Datos* (CMBD), un sistema comúnmente utilizado en los hospitales de Cataluña y otros centros hospitalarios de España. Este sistema de notificación tiene como objetivo unificar la información sobre morbilidad, tratamientos y otros aspectos clínicos, permitiendo una evaluación precisa de los servicios sanitarios prestados y facilitando comparaciones entre centros hospitalarios [19].

Es importante destacar que los datos utilizados para este proyecto fueron completamente anonimizados antes de su análisis. Todos los identificadores personales de los pacientes, como nombres, números de identificación y cualquier otra información que pudiera permitir la identificación directa o indirecta de los individuos, fueron eliminados o permutados. Este proceso de anonimización se llevó a cabo en cumplimiento con las normativas vigentes de protección de datos personales, garantizando que la privacidad de los pacientes fuera estrictamente protegida.

Además, el proyecto se desarrolló bajo un acuerdo de confidencialidad firmado entre las partes involucradas, asegurando el manejo adecuado y seguro de los datos sensibles. El proyecto también contó con la aprobación del Comité de Ética de Investigación con Medicamentos (CEIm), cumpliendo con todos los requisitos éticos y legales necesarios para el tratamiento de datos de salud. Esta aprobación incluyó la validación del protocolo de anonimización y la verificación de que los datos procesados no contenían información que pudiera comprometer la identidad de los pacientes.

5.1.1. Estructura de los datos y variables

Las variables contenidas en el CMBD incluyen información identificativa y demográfica de los pacientes, como el *CIP* (Código de Identificación Personal), *Data_naix* (fecha de nacimiento), *Sexe* (sexo) y lugar de residencia. También incluyen información detallada sobre el proceso de ingreso y alta hospitalaria, tales como la *Data_ingres* (fecha de ingreso), *Data_alta* (fecha de alta), y las circunstancias de ingreso y alta.

Entre las variables más relevantes para el proyecto se encontraban las relacionadas con los diagnósticos y los procedimientos. Estas variables se codifican siguiendo la *Clasificación Internacional de Enfermedades* (CIM-10), lo que permite la normalización y comparabilidad de los datos. Los diagnósticos se identifican mediante las variables *DP* (diagnóstico principal) y *DS1*...*DS14* (diagnósticos secundarios), mientras que los procedimientos se describen en las variables *PP* (procedimiento principal) y *PS1*...*PS14* (procedimientos secundarios).

Cada diagnóstico se combina con su indicador POA correspondiente, agregando un sufijo al código del diagnóstico (por ejemplo, `-S` para 'Sí', `-N` para 'No', `-D` para Desconocido, `-I` para Indeterminado Clínicamente, `-E` para Exempt). Esto permite diferenciar entre diagnósticos presentes al ingreso y los adquiridos durante la estancia hospitalaria.

El *Manual de Notificación de Hospitales Generales de Agudos* del CMBD [19] detalla las normas de codificación y uso de estas variables, asegurando que la información sea consistente en todos los hospitales que usan este sistema.

5.2 Preparación de Datos para Clustering

Una vez identificadas y transformadas las variables relevantes, el siguiente paso fue preparar los datos para la aplicación de algoritmos de *clustering*. Dado que el conjunto de datos original estaba compuesto por episodios clínicos y no por pacientes individuales, se procedió a agrupar episodios que presumiblemente correspondían al mismo paciente. Para ello, se asignaron al mismo paciente aquellos episodios que tenían la misma fecha de nacimiento y sexo. Esta estrategia se adoptó porque el algoritmo de *clustering* empleado fue diseñado y validado para agrupar pacientes, no episodios. Al unificar los episodios bajo supuestos pacientes con características demográficas idénticas, adaptamos el conjunto de datos a los requisitos del algoritmo y aseguramos la coherencia en el análisis.

Este proceso también incluyó la normalización de los datos para asegurar que todas las variables tuvieran un rango comparable, lo cual es fundamental para que los algoritmos de *clustering* funcionen de manera efectiva.

Además, se determinó el número óptimo de *clusters* a formar. Este parámetro se definió en función de los objetivos del análisis, permitiendo ajustar la granularidad de los resultados según las necesidades específicas del proyecto. La preparación de los datos también contempló la eliminación de valores atípicos y la imputación de valores faltantes, asegurando que el conjunto de datos estuviera limpio y listo para su procesamiento.

5.2.1. Descripción del Algoritmo de Clustering

El algoritmo de *clustering* utilizado en este proyecto clasifica a los pacientes en diferentes grupos (*clusters*) basados en sus características clínicas, maximizando la homogeneidad dentro de cada *cluster* y la heterogeneidad entre ellos. Este algoritmo emplea un modelo de Naive Bayes con variables binarias y se basa en el método de los momentos, lo que permite calcular las correlaciones entre las variables observadas y descubrir patrones en los datos sin necesidad de definir una función de distancia, como ocurre en otros métodos más tradicionales como el k-means [3].

Es importante no confundir este algoritmo con la técnica de construcción de predictores lineales también llamada Naive Bayes. Aunque ambas técnicas aplican la regla de Bayes, son muy diferentes: una es supervisada (predicción) y la otra no supervisada. Además, el algoritmo de clustering basado en Naive Bayes es matemáticamente mucho más complejo que el comparativamente simple algoritmo de predicción.

En lugar de utilizar exclusivamente métodos basados en la distancia, este enfoque se fundamenta en modelos de mezcla que describen los datos observados como resultados de distribuciones conjuntas, lo que permite obtener interpretaciones más naturales y objetivas de los datos clínicos. Para mejorar la eficiencia, el algoritmo utiliza la descomposición de tensores para descomponer las correlaciones entre las variables, seguido por una optimización mediante el algoritmo de *Expectation Maximization* (EM), lo que garantiza una convergencia rápida y resultados óptimos [3].

Este enfoque permite trabajar con grandes volúmenes de datos hospitalarios y obtener grupos de pacientes con características similares, lo que facilita el análisis posterior y la identificación de patrones clínicos relevantes. Aunque el algoritmo no es el foco principal de este trabajo, su aplicación es fundamental para agrupar pacientes y realizar análisis en base a los *clusters* generados.

5.2.2. Integración de los Resultados del Clustering y análisis

Tras completar el *clustering*, los resultados se integraron en el conjunto de datos original, etiquetando cada registro de paciente con su *cluster* correspondiente. Esto permitió mantener todas las variables originales junto con la nueva clasificación, facilitando así análisis más profundos y específicos. La incorporación de esta información permitió explorar patrones y características comunes dentro de cada *cluster*, lo que proporcionó una base sólida para la posterior generación de datos sintéticos.

5.3 Cálculo de métricas por *cluster*

Para analizar las características de cada *cluster*, se calculan una serie de métricas que permiten comprender mejor la composición y distribución de los datos dentro de cada grupo. A continuación, se detalla el proceso realizado para calcular estas métricas para cada *cluster*:

5.3.1. Frecuencia del *cluster*

Se calcula la frecuencia de cada *cluster* como la proporción de registros que pertenecen a dicho *cluster* en relación con el total de registros del conjunto de datos. Esta medida indica el peso relativo de cada *cluster* y permite identificar cuáles son los grupos más representativos en el conjunto de datos.

5.3.2. Distribución de episodios por paciente

Para entender cómo se distribuyen los episodios entre los pacientes dentro de cada *cluster*, se analiza la cantidad de episodios que tiene cada paciente en ese grupo. Se cuenta el número de episodios asociados a cada paciente y se calcula la frecuencia relativa de pacientes para cada posible número de episodios. Esto permite conocer si en el *cluster* predominan pacientes con muchos episodios o con pocos, proporcionando información sobre patrones de hospitalización.

5.3.3. Variables categóricas

Para cada variable categórica relevante, como *Sexe*, *Circ_admiss* o *Procedencia ingres*, se calcula la frecuencia relativa de cada uno de sus valores posibles dentro del *cluster*, incluyendo la proporción de valores nulos o faltantes. Esto se realiza contando el número de ocurrencias de cada valor y normalizando respecto al total de registros del *cluster*. De esta manera, se obtiene la distribución de estas variables en cada grupo, lo que es fundamental para mantener la coherencia al generar datos sintéticos.

5.3.4. Variables continuas

Para las variables continuas, como *Edat* (edad), *Dies_estada* (días de estancia) y *DRG*, se emplea una técnica de agrupamiento en intervalos (*binning*) para analizar su distribución dentro del *cluster*. El rango de valores de cada variable se divide en un número determinado de intervalos, adaptando el número de *bins* en función de la dispersión de los datos (por ejemplo, se puede aumentar el número de *bins* si la variable tiene un rango muy amplio). Luego, se calcula la proporción de registros que caen dentro de cada intervalo, obteniendo así una distribución que refleja la variabilidad de la variable en el *cluster* y que facilita la generación de valores sintéticos coherentes.

5.3.5. Diagnósticos y procedimientos más frecuentes

Se identifican los diagnósticos y procedimientos más frecuentes en cada *cluster*, tanto principales como secundarios, teniendo en cuenta también los indicadores de presencia al ingreso (POA, *Present on Admission*). El proceso realizado es el siguiente:

- **Procesamiento de diagnósticos:** Se analizan el diagnóstico principal (*DP*) y los diagnósticos secundarios (*DS1* a *DS14*). Cada diagnóstico se combina con su indicador POA correspondiente, agregando un sufijo al código del diagnóstico (por ejemplo, `-S` para 'Sí', `-N` para 'No', `-D` para Desconocido, `-I` para Indeterminado Clínicamente, `-E` para Exempt). Esto permite diferenciar entre diagnósticos presentes al ingreso y los adquiridos durante la estancia hospitalaria.
- **Cálculo de frecuencias:** Se cuentan las ocurrencias de cada diagnóstico combinado con su indicador POA dentro del *cluster* y se normalizan las frecuencias para obtener

proporciones relativas. Se calculan por separado las frecuencias para el diagnóstico principal y para los diagnósticos secundarios.

- **Cálculo de valores faltantes:** Se calcula la proporción de registros con diagnósticos faltantes o nulos, tanto para el diagnóstico principal como para los secundarios. Esto incluye tanto valores *NaN* como cadenas vacías, para asegurar una contabilización precisa de los datos faltantes.
- **Procesamiento de procedimientos:** De manera análoga, se analizan el procedimiento principal (*PP*) y los procedimientos secundarios (*PS1* a *PS14*). Se cuentan las ocurrencias de cada código de procedimiento y se normalizan las frecuencias relativas.
- **Cálculo de valores faltantes en procedimientos:** Se calcula la proporción de registros con procedimientos faltantes o nulos, tanto para el procedimiento principal como para los secundarios.

Estas métricas proporcionan información detallada sobre las condiciones clínicas y las intervenciones más comunes en cada *cluster*, así como sobre la completitud de los datos. Esto es esencial para generar datos sintéticos que reflejen fielmente las características de los datos reales y para asegurar que los diagnósticos y procedimientos asignados a los episodios sintéticos sean coherentes con la práctica clínica observada.

5.3.6. Resumen de métricas por *cluster*

Al finalizar el cálculo, se obtiene para cada *cluster* un conjunto completo de métricas que incluye:

- **Frecuencia del *cluster*:** Proporción de registros pertenecientes al *cluster* respecto al total.
- **Distribución de episodios por paciente:** Frecuencia relativa de pacientes según la cantidad de episodios que presentan en el *cluster*.
- **Variables categóricas:** Distribuciones de frecuencias relativas para cada valor de las variables categóricas, incluyendo valores faltantes.
- **Variables continuas:** Distribuciones de frecuencias en intervalos definidos para cada variable continua, reflejando su variabilidad dentro del *cluster*.
- **Diagnósticos y procedimientos:** Listas de los diagnósticos y procedimientos más frecuentes (tanto principales como secundarios), con sus respectivas frecuencias relativas y proporciones de valores faltantes.

Estas métricas son fundamentales para el posterior proceso de generación de datos sintéticos, ya que proporcionan la base estadística necesaria para replicar las características de los datos originales en cada *cluster*. Al utilizar estas métricas, se garantiza que los datos sintéticos generados mantengan la coherencia y representatividad de los patrones clínicos y administrativos observados en los datos reales.

5.4 Generación de datos sintéticos

Con las distribuciones de las variables definidas para cada *cluster*, se procedió a la generación de datos sintéticos. El objetivo fue crear registros que mantuvieran las características y

distribuciones observadas en los datos originales, garantizando su coherencia y realismo. Para ello, se implementaron funciones que replican las distribuciones previamente calculadas, asegurando que los valores generados reflejaran adecuadamente la variabilidad y complejidad de los datos clínicos reales.

El proceso de generación se inició con la asignación de un ID único para cada paciente, utilizando una función de codificación *base64* para asegurar la unicidad. Se determinó el número de episodios que tendría cada paciente, basado en la distribución de episodios por paciente en el *cluster*, replicando así la frecuencia de hospitalizaciones observada en los datos reales.

Para las variables categóricas, como *Sexe*, *Circ admiss* y *Procedencia ingres*, se utilizaron las distribuciones de frecuencias relativas obtenidas en el análisis de métricas del *cluster* correspondiente. Esto permitió asignar valores a los pacientes de manera proporcional a su ocurrencia en los datos originales, manteniendo la coherencia con las características observadas.

En el caso de las variables continuas, como *Edat* (edad), *Dies estada* (días de estancia) y *DRG*, se aplicaron técnicas de *binning* para dividir el rango de valores en intervalos. Al generar los datos sintéticos, se seleccionó un intervalo basado en la distribución de frecuencias y luego se asignó un valor aleatorio dentro de ese intervalo. Esto garantizó que los valores generados reflejaran la distribución original de estas variables, preservando la variabilidad y evitando valores atípicos no representativos.

Los diagnósticos y procedimientos fueron asignados utilizando las distribuciones de frecuencias calculadas para cada *cluster*, incluyendo tanto el diagnóstico principal (*DP*) y el procedimiento principal (*PP*), como los diagnósticos y procedimientos secundarios (*DS1* a *DS14* y *PS1* a *PS14*). Se consideraron también las probabilidades de valores faltantes (*missing*) para estas variables, permitiendo que en algunos casos no se asignaran diagnósticos o procedimientos, reflejando así la realidad de los datos originales donde no siempre se dispone de información completa.

Además, se prestó especial atención a los indicadores de presencia al ingreso (POA, *Present on Admission*). Los diagnósticos fueron procesados para extraer el indicador POA y limpiar los sufijos añadidos durante el análisis de métricas. Esto permitió mantener la coherencia en la asignación de diagnósticos y asegurar que los episodios sintéticos reflejaran correctamente la situación clínica de los pacientes al momento del ingreso.

El proceso general de generación de episodios sintéticos se resumió en los siguientes pasos:

1. **Asignación de pacientes y episodios:** Se generó un ID único para cada paciente y se determinó el número de episodios asociados, basándose en la distribución de episodios por paciente del *cluster*. Esto replicó la frecuencia de hospitalizaciones observada en los datos reales.
2. **Asignación de variables demográficas y temporales:** Para cada episodio, se asignaron variables como edad, sexo, fecha de nacimiento, fechas de ingreso y alta, y otras variables administrativas, utilizando las distribuciones correspondientes del *cluster*. Se

garantizó la coherencia temporal, asegurando que la fecha de alta fuera posterior a la de ingreso y que la edad correspondiera con la fecha de nacimiento.

3. **Asignación de diagnósticos y procedimientos:** Se asignaron los diagnósticos principales y secundarios, así como los procedimientos, considerando las distribuciones de frecuencias y las probabilidades de valores faltantes. Se incluyeron los indicadores POA, diferenciando entre diagnósticos presentes al ingreso y aquellos adquiridos durante la estancia hospitalaria.

4. **Construcción de registros sintéticos:** Se compiló toda la información generada en registros únicos que representan episodios sintéticos, incluyendo todas las variables relevantes y el identificador del *cluster* al que pertenece. Esto mantuvo la estructura original de los datos y facilitó su análisis posterior.

5. **Generación del conjunto completo de datos sintéticos:** Se repitieron los pasos anteriores para crear el número total de episodios sintéticos deseados, distribuidos entre los *clusters* según sus frecuencias relativas. Así, la distribución general de los datos sintéticos reflejó la composición original de los *clusters* identificados en el análisis.

Al finalizar, se obtuvo un conjunto de datos sintéticos que refleja las características y distribuciones de los datos originales, preservando la coherencia y realismo necesarios para su uso en análisis posteriores. Este enfoque permitió generar datos clínicos sintéticos de alta calidad, manteniendo la privacidad de los pacientes y facilitando su aplicación en estudios, desarrollo de modelos predictivos y otras investigaciones en el ámbito de la salud.

5.5 Validación de Datos Sintéticos

Para validar la calidad de los datos sintéticos generados, se utilizó un modelo de Random Forest que comparó estos datos con los datos reales. El objetivo de la validación fue evaluar si el modelo podía distinguir entre los datos reales y los sintéticos; una mayor dificultad para realizar esta distinción indicaría una alta calidad en los datos sintéticos.

El proceso de validación se inició cargando los conjuntos de datos sintéticos y reales. Se seleccionaron muestras aleatorias de ambos conjuntos, asegurando que cada muestra contuviera 1000 registros para mantener una comparación balanceada. A cada registro se le asignó una etiqueta binaria que indicaba si provenía del conjunto sintético o real. Posteriormente, se identificaron las columnas comunes entre ambos conjuntos de datos para asegurar que el análisis se realizara sobre variables comparables.

Antes de entrenar el modelo, se realizó una preprocesamiento de los datos que incluyó la agrupación de las columnas de diagnósticos (*DS1* a *DS14*) en una sola columna denominada *other_diagnostics*, y de las columnas de tratamientos (*PS1* a *PS14*) en otra columna denominada *other_treatments*. Además, se eliminaron columnas no necesarias para el análisis, como *Id_pacient*, las fechas de nacimiento, ingreso y alta, y cualquier otra columna irrelevante. Las columnas relacionadas con los indicadores de presencia al ingreso (*POA*) también fueron eliminadas para enfocarse en las variables clínicas y administrativas principales.

Para manejar los valores faltantes, se imputaron con un valor placeholder genérico (`-1`), asegurando que el modelo pudiera procesar los datos sin interrupciones. Las variables cate-

góricas fueron transformadas mediante codificación *one-hot*, lo que permitió al modelo de Random Forest evaluar de manera eficiente tanto variables continuas como categóricas.

Una vez preprocesados los datos, se definieron las características (*features*) y la variable objetivo (*target*) que indicaba si el registro era real o sintético. Los datos fueron divididos en conjuntos de entrenamiento y prueba, utilizando un 70 % para entrenamiento y un 30 % para prueba, con una semilla aleatoria para garantizar la reproducibilidad.

El modelo de Random Forest fue inicializado con una semilla aleatoria para asegurar consistencia en los resultados y fue entrenado utilizando el conjunto de entrenamiento. Posteriormente, se realizaron predicciones sobre el conjunto de prueba para evaluar la capacidad del modelo para distinguir entre datos reales y sintéticos.

La evaluación del modelo se realizó mediante la generación de una matriz de confusión y un informe de clasificación, que incluyeron métricas como precisión, recall, *f1-score* y exactitud (*accuracy*). Estas métricas permitieron cuantificar la capacidad del modelo para identificar correctamente la fuente de los datos. Un rendimiento bajo en la distinción entre datos reales y sintéticos, cerca del 50 porciento, indicaría que los datos sintéticos son altamente realistas y difíciles de diferenciar de los datos reales.

Además, se analizó la importancia de las características utilizadas por el modelo para determinar cuáles variables influían más en la capacidad de distinción. Las 20 características más importantes fueron visualizadas gráficamente mediante gráficos de barras, lo que permitió identificar patrones clave que diferenciaban ambos conjuntos de datos. Este análisis ayudó a comprender qué variables tenían mayor impacto en la diferenciación, proporcionando información valiosa sobre la similitud y diferencias entre los datos reales y sintéticos.

Finalmente, se compararon las distribuciones de las variables continuas y categóricas entre los datos reales y sintéticos. Para las variables continuas, como *Dies_estada* y *Edat*, se realizaron gráficas de densidad (*density plots*) que permitieron visualizar la similitud en la distribución de estos valores entre ambos conjuntos. En el caso de las variables categóricas, se utilizaron gráficos de barras para comparar las frecuencias de cada valor, asegurando que las distribuciones en los datos sintéticos reflejaran adecuadamente las observadas en los datos reales.

5.6 Estructura del código y Ajuste de Parámetros

La figura a continuación ilustra la estructura del código utilizado para realizar este proceso de ajuste de parámetros, integrando todas las fases descritas anteriormente. El código se divide en tres grandes bloques: procesamiento y clusterización, cálculo de métricas por clusters, y generación de datos sintéticos, como se indica en los recuadros con líneas intercontinuas.

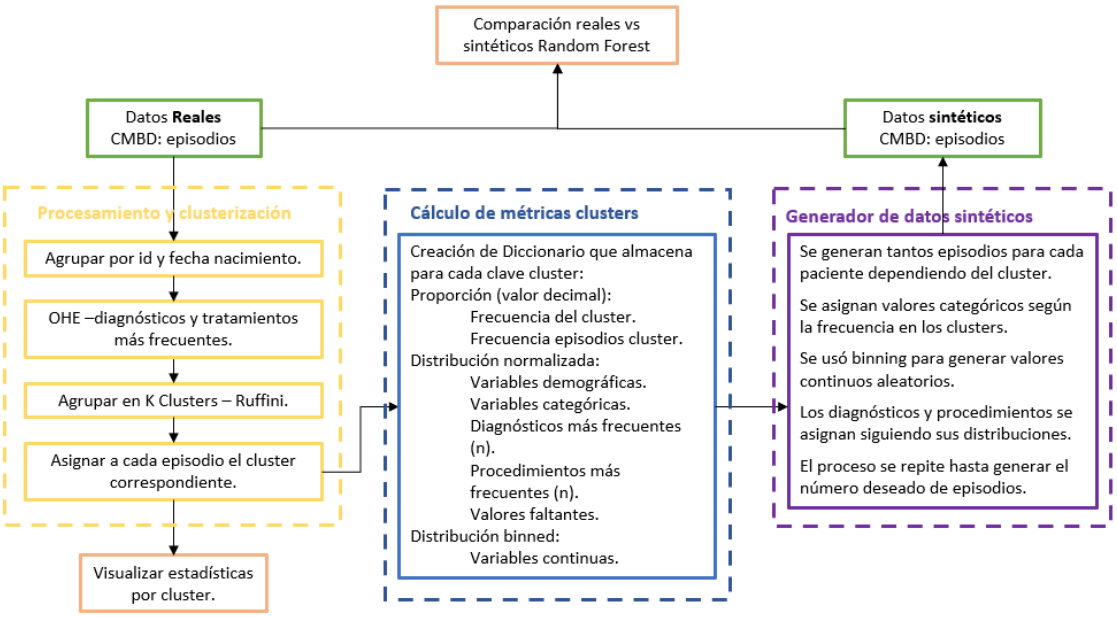

Figura 1. Estructura del Código para el Ajuste de Parámetros y Generación de Datos Sintéticos.

Tal y como observamos en el diagrama, durante el desarrollo del proceso de generación de datos sintéticos se hace uso de tres parámetros clave: *frecuencia*, *k* y *n*. La optimización de estos parámetros fue esencial para asegurar que los datos sintéticos replicaran fielmente las características de los datos reales y proporcionaran resultados de alta calidad al ser evaluados por el modelo de validación.

- **Frecuencia:** Este parámetro controla el umbral mínimo para considerar diagnósticos y tratamientos significativos. Al variar el valor de *frecuencia*, se ajusta qué proporción de diagnósticos y tratamientos se incluye en la generación de datos sintéticos. Valores más bajos de *frecuencia* permiten incluir diagnósticos menos comunes, mientras que valores más altos limitan la generación a los más frecuentes.
- k**:** Este parámetro define el número de *clusters* en los que se agrupan los pacientes durante el proceso de *clustering*. El valor de k afecta la granularidad de los grupos generados y, por ende, la representación de las características clínicas dentro de cada *cluster*. Un valor más alto de k crea más grupos pequeños, mientras que un valor más bajo genera menos grupos, pero más grandes.
- n**:** Este parámetro establece el número de diagnósticos y procedimientos más comunes que se utilizan para caracterizar cada *cluster*. Al ajustar n, se controla el nivel de detalle en la representación de las condiciones médicas y los tratamientos. Valores más altos de n permiten una mayor inclusión de diagnósticos y procedimientos secundarios, mientras que valores más bajos se centran en los principales.

El proceso de ajuste de estos parámetros se realizó mediante una búsqueda en cuadrícula (*grid search*), en la que se probaron diferentes combinaciones de *frecuencia*, *k* y *n* para encontrar la configuración óptima. Para cada combinación, se generaron datos sintéticos y se evaluaron utilizando un modelo de Random Forest, que intentaba distinguir entre los da-

tos sintéticos y los reales. Las métricas utilizadas en la evaluación incluyeron la precisión (*precision*), el *recall*, el *f1-score* y la exactitud (*accuracy*) del modelo.

Cada iteración de la búsqueda consistió en los siguientes pasos:

Algorithm 1: Iteraciones para Ajuste de Parámetros

Data: Rangos de `frequencia`, `k`, y `n`
foreach *frequencia en rango_frequencia* **do**
  Realizar el preprocesamiento de los datos Filtrar diagnósticos y tratamientos más frecuentes;
  foreach *k en rango_k* **do**
    Aplicar Clustering (Naive Bayes) Generar `k` clusters de los pacientes;
    foreach *n en rango_n* **do**
      Calcular métricas por cluster Calcular frecuencias de diagnósticos y tratamientos;
      Generar datos sintéticos Usar métricas de los clusters para crear datos sintéticos de episodios clínicos;
      Validar la calidad de los datos sintéticos Evaluar la capacidad del modelo para distinguir entre datos reales y sintéticos;
      Comparar los datos reales vs sintéticos;
      Evaluar el rendimiento de la iteración Actualizar los mejores resultados y parámetros óptimos;

El resultado final de este proceso fue la selección de la mejor combinación de los parámetros *frecuencia*, *k* y *n*, que maximizó el *f1-score* del modelo, reflejando una alta calidad en los datos sintéticos generados. Esta combinación se seleccionó como la configuración óptima y se utilizó para generar el conjunto final de datos sintéticos.

Este enfoque permitió identificar la configuración que mejor equilibraba la precisión y la representatividad de los datos sintéticos, garantizando su utilidad en análisis futuros y validando su similitud con los datos reales.

5.7 Recursos requeridos

A continuación, se enumeran los recursos utilizados para la ejecución del proyecto, organizados en diferentes categorías según su naturaleza.

Software y librerías

- **Python:** Lenguaje de programación utilizado para el desarrollo del algoritmo de *clustering*, análisis de datos y generación de datos sintéticos.
- **Pandas:** Librería esencial para la manipulación y análisis de datos tabulares. Se utilizó para cargar, procesar, limpiar y transformar los conjuntos de datos.
- **NumPy:** Librería fundamental para operaciones numéricas, manipulación de arreglos y cálculos matemáticos, utilizada en las funciones de *clustering* y generación de datos sintéticos.
- **Scikit-learn:** Librería de *machine learning* utilizada para implementar el modelo de Random Forest para la validación de datos sintéticos y el *train-test split* para la evaluación del modelo. También se usó para el preprocesamiento de los datos, como la binarización de múltiples etiquetas con *MultiLabelBinarizer*.

- **Faker:** Librería para generar datos ficticios, utilizada en este proyecto para crear valores sintéticos como fechas de nacimiento y admisión hospitalaria en el conjunto de datos generado.
- **Base64:** Librería utilizada para generar identificadores únicos codificados en *base64* para los pacientes sintéticos.
- **Datetime:** Módulo de Python empleado para la manipulación de fechas y tiempos, clave en la generación de eventos temporales en los episodios sintéticos.
- **OS:** Módulo utilizado para la manipulación de archivos del sistema operativo, como la lectura y escritura de archivos de datos.
- **Time:** Módulo de Python empleado para medir tiempos de ejecución y optimizar el proceso de generación de datos sintéticos.

Recursos de computación

- **Ordenador personal:** El desarrollo, pruebas y ejecución del software se realizaron en un equipo con las siguientes especificaciones:
  - Procesador: Intel(R) Core(TM) i7-11800H @ 2.30GHz.
  - Memoria RAM: 32 GB.
  - Sistema operativo: Windows 10 Pro (versión 22H2).

  Estas especificaciones aseguran que el equipo tenga suficiente capacidad de procesamiento y memoria para manejar el análisis de datos y el entrenamiento de modelos de *machine learning* utilizados en el proyecto.

5.8 Presupuesto

En esta sección se detalla el presupuesto estimado para el desarrollo del proyecto. Cabe mencionar que, debido al uso de software gratuito y recursos accesibles a través de internet, no se incurrieron en gastos adicionales significativos más allá del tiempo dedicado y el equipo utilizado. El equipo técnico pertenece a la infraestructura existente y no fue adquirido específicamente para este proyecto.

Tipo de coste	Valor	Comentarios
Horas de trabajo en el proyecto	150 horas	Las horas son aproximadas y corresponden a la dedicación necesaria para el desarrollo del código, análisis de datos y redacción del informe.
Equipo técnico utilizado	1.000 €	El proyecto se realizó utilizando un equipo informático con procesador Intel i7 y 32 GB de RAM. El coste del equipo es estimado y se prorratea considerando su uso en este y otros proyectos.
Software utilizado	0 €	Todo el software empleado es de código abierto y gratuito, accesible a través de internet.
Estudios e informes	0 €	No se requirió la contratación de estudios externos ni la intervención de consultorías.
Materiales empleados	0 €	No se emplearon materiales adicionales más allá del equipo técnico mencionado.

Tabla 2. Presupuesto del proyecto

El coste total del proyecto, como se observa en la Tabla 2, es mínimo debido al uso de software gratuito y al aprovechamiento de infraestructura existente.

5.9 Viabilidad

Este apartado evalúa brevemente la viabilidad económica y la sostenibilidad del proyecto a largo plazo, considerando los costos y beneficios asociados.

5.9.1. Análisis de viabilidad económica

El desarrollo del proyecto se ha realizado con un costo económico bajo, utilizando herramientas de software libre como Python. El principal costo está relacionado con el tiempo invertido, estimado en 100 horas de trabajo de un desarrollador.

Los beneficios potenciales, como la capacidad de generar datos sintéticos para investigación médica sin comprometer la privacidad de los pacientes, superan los costos. La metodología desarrollada puede contribuir a mejorar la calidad de los tratamientos y reducir errores clínicos, lo que implica un impacto positivo en términos económicos y sociales.

5.9.2. Sostenibilidad a futuro

El proyecto es sostenible en su fase actual con recursos mínimos. No obstante, para una implementación a mayor escala se requieren mejoras en tres áreas clave:

- **Federación del sistema:** Será necesario integrar el algoritmo en sistemas de datos distribuidos a nivel hospitalario, cumpliendo con las normativas de seguridad y privacidad.
- **Mantenimiento:** El software debe ser actualizado regularmente para adaptarse a nuevas normativas y avances tecnológicos.
- **Capacitación:** Se debe capacitar a los usuarios y ofrecer soporte técnico continuo para facilitar su adopción y uso.

Aunque estas mejoras implican costos adicionales, el beneficio potencial justifica la inversión.

5.9.3. Recomendaciones para la sostenibilidad

Para asegurar la sostenibilidad a largo plazo, se proponen las siguientes acciones:

- **Alianzas estratégicas:** Colaborar con instituciones médicas y universidades para asegurar financiamiento y acceso a infraestructura.
- **Modelo de negocio:** Explorar la creación de un modelo de licencias o acuerdos de colaboración con empresas del sector salud.
- **Financiación pública:** Solicitar subvenciones gubernamentales para apoyar el desarrollo y la implementación del proyecto.

En resumen, el proyecto es económicamente viable y tiene un impacto potencial significativo en el ámbito de la salud, siempre que se implementen medidas para asegurar su sostenibilidad a largo plazo.

Capítulo 6. Resultados

En esta sección se presentan los resultados obtenidos a lo largo del proyecto, detallando el plan de pruebas realizado y los análisis métricos que permiten evaluar la eficacia y validez del método propuesto.

Plan de Pruebas

El plan de pruebas desarrollado para validar el proyecto incluyó los siguientes pasos clave:

1. **Validación de la Integración de Datos:** Se verificó que el algoritmo fuera capaz de procesar correctamente los datos encriptados de entrada y producir salidas coherentes y estructuradas, alineadas con los requisitos iniciales del proyecto.
2. **Pruebas de Robustez:** Se evaluó la estabilidad y escalabilidad del algoritmo bajo diferentes configuraciones de parámetros y volúmenes de datos. Las pruebas demostraron que el algoritmo es robusto y puede manejar conjuntos de datos de tamaño variable sin pérdida significativa de precisión o eficiencia.
3. **Evaluación Comparativa:** Se compararon los datos sintéticos generados con los datos reales utilizando un modelo de Random Forest. El objetivo fue medir la capacidad del modelo para distinguir entre los dos conjuntos de datos. Los resultados indicaron que el modelo tuvo dificultades para diferenciar entre los datos reales y los sintéticos, lo que confirma la calidad y utilidad del conjunto de datos sintéticos.

6.1 Métricas de los Episodios por Cluster

Para verificar la correcta clusterización de los datos, se analizaron diversas métricas dentro de cada *cluster*. Estas métricas incluyeron la frecuencia de los *clusters*, la distribución de episodios por paciente, y la distribución de variables categóricas y continuas. Al evaluar estas métricas, se pudo confirmar que los *clusters* generados reflejan adecuadamente las características observadas en los datos originales, asegurando que cada grupo de pacientes presenta patrones consistentes y diferenciables.

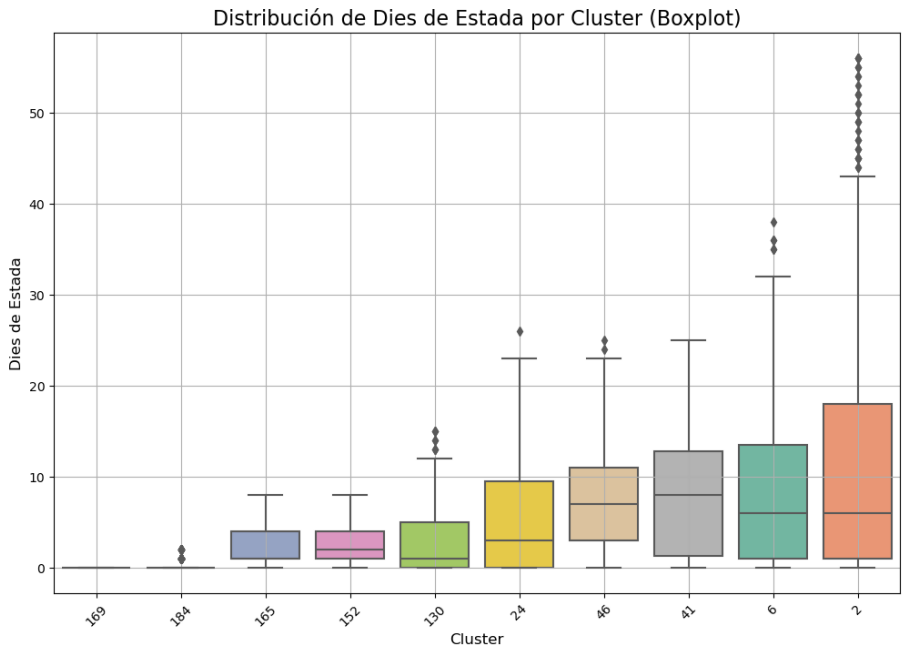

Figura 2. Distribución de los días de estancia (*Dies de Estada*) por *cluster*.

El gráfico en la Figura 2 muestra la distribución de los días de estancia (*Dies de Estada*) para cada *cluster*. Se puede observar que algunos *clusters* presentan una mayor dispersión, como es el caso del *cluster* 2, lo que indica que este grupo de pacientes tiene una estancia hospitalaria más variable. En contraste, otros *clusters* como el 169 o el 184 muestran una distribución más compacta, reflejando una estancia hospitalaria más homogénea.

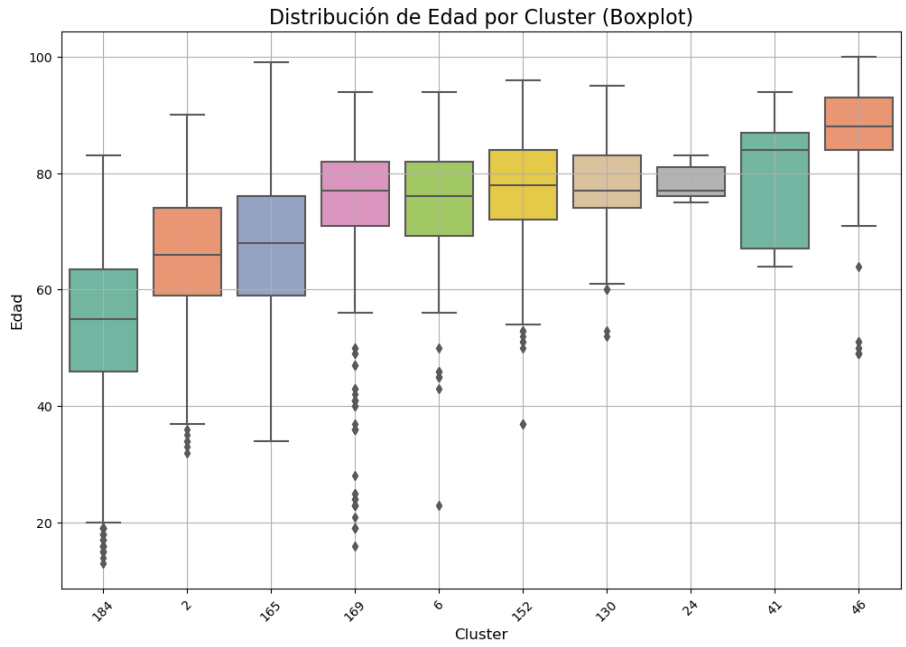

Figura 3. Distribución de la edad (*Edat*) por *cluster*.

En la Figura 3, se ilustra la distribución de la edad para los diferentes *clusters*. Los *clusters* 6 y 2 destacan por agrupar a una mayor proporción de pacientes de edad avanzada, mientras que otros *clusters*, como el 184 y el 165, muestran una distribución más equilibrada entre pacientes de distintas edades.

En conjunto, estos resultados permiten confirmar que la segmentación por *clusters* captura adecuadamente la variabilidad en variables clave como los días de estancia y la edad, proporcionando una base sólida para realizar análisis más detallados sobre cada grupo de pacientes.

6.2 Extracción de Métricas por Cluster

Se realizaron diversas visualizaciones para comparar las métricas extraídas del *dataset* con los datos originales de cada *cluster*. Estas visualizaciones permitieron verificar la correcta extracción de las métricas y asegurarse de que reflejaran las características reales de los datos. En particular, se analizaron distribuciones de variables continuas, como la edad y los días de estancia, junto con la frecuencia de diagnósticos y tratamientos.

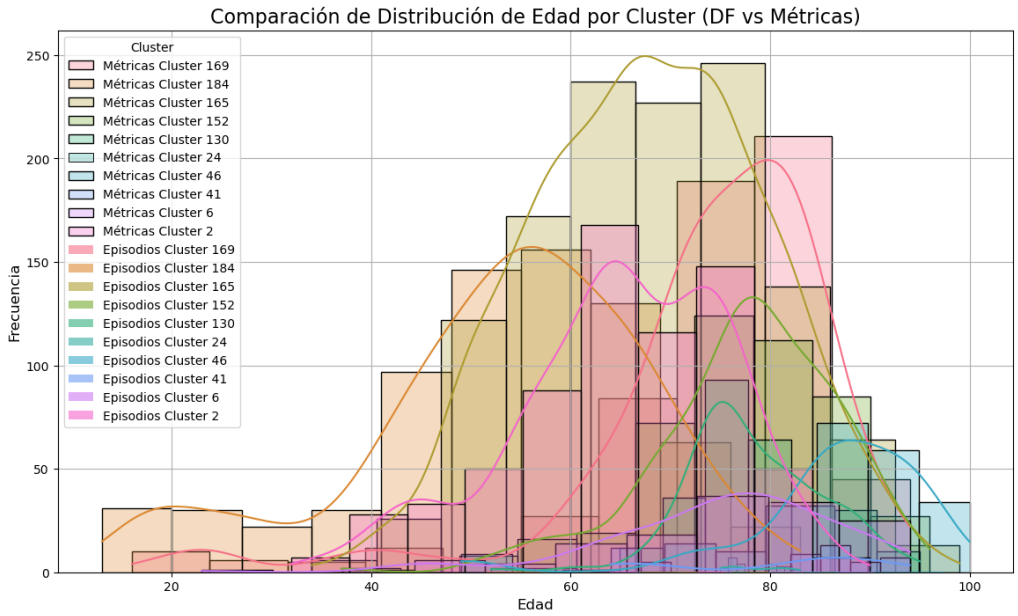

Figura 4. Comparación de la distribución de edad entre los datos originales y las métricas extraídas por *cluster*.

En la Figura 4 se muestra la comparación de las distribuciones de edad por *cluster*. Esta visualización confirma que las métricas extraídas del *dataset* de los *clusters* corresponden correctamente con las distribuciones observadas en los datos originales. Las similitudes en las distribuciones validan la coherencia de las métricas, asegurando su validez para análisis posteriores.

A continuación, se presenta la comparación de la distribución de los días de estancia entre los datos originales y las métricas extraídas por *cluster*.

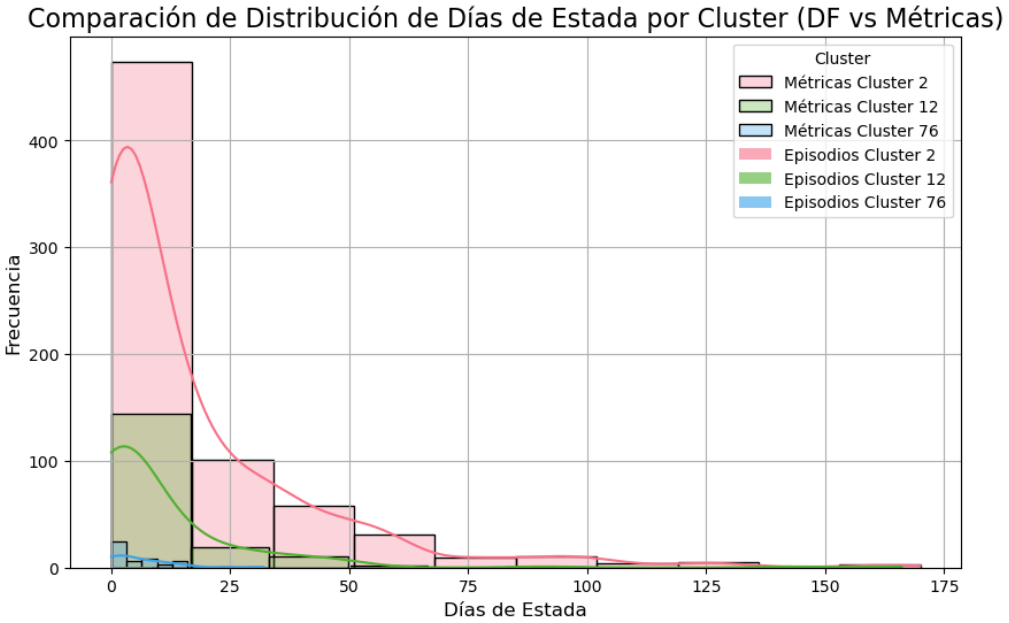

Figura 5. Comparación de la distribución de días de estancia entre los datos originales y las métricas extraídas por *cluster*.

En la Figura 5, se comparan las distribuciones de los días de estancia por *cluster*. La visualización revela que las métricas extraídas logran capturar con precisión las características de los datos originales. Cada *cluster* muestra una distribución distinta, y las métricas calculadas reflejan con fidelidad la forma y dispersión de los datos originales, incluyendo la variabilidad en las estancias cortas y largas.

Esta coherencia entre las distribuciones de los datos originales y las métricas validadas es esencial para asegurar que los análisis posteriores basados en estas métricas sean fiables y representativos de la realidad. La correcta extracción y representación de las métricas, como la duración de la estancia hospitalaria, garantiza que los patrones y tendencias observadas en los datos reales se mantengan, permitiendo un análisis clínico y estadístico adecuado.

A continuación se presenta un ejemplo de las métricas extraídas para el *cluster* 165:

Cluster 165:

frecuencia_cluster: 0.0184

episode_distribution: {1: 0.425, 2: 0.2867, 3: 0.1683, 4: 0.07}

Sexe: {0: 0.7250, 1: 0.2750}

Circ_admiss: {1: 0.5228, 2: 0.4772}

Procedencia_ingres: {8: 0.8953, 7: 0.0568, 2: 0.0288, 1: 0.0176}

Circ_alta: {1: 0.8753, 2: 0.0592, 9: 0.0568, 6: 0.0048}

Servei_alta: {Cardiologia: 0.4668, Urgències: 0.1207, Oftalmologia: 0.0855, Cirurgia_general: 0.0711}

Edat: {Interval(56.75, 60.0, closed='right'): 0.1055, Interval(60.0, 63.25, closed='right'): 0.0791, Interval(63.25, 66.5, closed='right'): 0.0847, Interval(66.5, 69.75, closed='right'): 0.1039}

Dies_estada: {Interval(-0.037, 1.8, closed='right'): 0.4972, Interval(1.8, 3.6, closed='right'): 0.2006, Interval(3.6, 5.4, closed='right'): 0.1399, Interval(5.4, 7.2, closed='right'): 0.0871}

Diagnósticos principales (DP): {Q6410-S-S: 0.1219, O09613-S-S: 0.1206, C531-S-S: 0.0747, T471X5A-S-S: 0.0577}
Otros diagnósticos: {K5752-S-S: 0.1509, L500-S-S: 0.1190, T8143XA-S-S: 0.0805, S82301B-S-S: 0.0672}

6.3 Búsqueda de Parámetros mediante Grid Search

El proceso de búsqueda de la mejor configuración de parámetros se llevó a cabo utilizando una estrategia de *Grid Search*. Se evaluaron diversas combinaciones de los parámetros clave: *frecuencia*, k (número de *clusters*), y n (número de diagnósticos), con el objetivo de identificar la configuración que mejorara la generación de datos sintéticos y maximizara su similitud con los datos reales.

Antes de presentar los resultados, es importante definir brevemente las métricas utilizadas para evaluar el desempeño del modelo:

- **Precisión (Precision):** Mide la proporción de verdaderos positivos sobre el total de predicciones positivas realizadas por el modelo. Indica la exactitud de las predicciones positivas.
- **Exhaustividad (Recall):** También conocida como sensibilidad, mide la proporción de verdaderos positivos identificados sobre el total de positivos reales en los datos. Indica la capacidad del modelo para encontrar todos los positivos.
- **Puntuación F1 (F1-Score):** Es la media armónica de la precisión y la exhaustividad. Proporciona un balance entre ambas métricas y es útil cuando existe un desequilibrio entre clases.
- **Exactitud (Accuracy):** Mide la proporción de predicciones correctas (tanto positivas como negativas) sobre el total de predicciones realizadas. En este contexto, una exactitud baja es deseable, ya que indica que el modelo tiene dificultades para distinguir entre datos reales y sintéticos.

A continuación se muestra una tabla que resume los resultados obtenidos para las distintas combinaciones de los parámetros *frecuencia*, k y n. La tabla incluye las métricas de precisión, exhaustividad, puntuación F1 y exactitud, así como el tiempo de ejecución para cada iteración.

Frecuencia	k clusters	n diagnostics	Precision	Recall	F1-Score	Accuracy	Tiempo de ejecución (s)
0.00005	50	20	0.8714	0.8683	0.8681	0.8683	287.75
0.00005	50	50	0.8719	0.8717	0.8717	0.8717	351.99
0.00005	50	200	0.8473	0.8467	0.8466	0.8467	419.75
0.00005	100	20	0.8743	0.8700	0.8697	0.8700	885.25
0.00005	100	50	0.8751	0.8750	0.8750	0.8750	948.78
0.00005	100	200	0.8419	0.8417	0.8417	0.8417	1017.94
0.00005	200	20	0.8479	0.8433	0.8429	0.8433	1772.53
0.00005	**200**	**50**	**0.8238**	**0.8200**	**0.8196**	**0.8200**	**1841.10**
0.00005	200	200	0.8406	0.8383	0.8380	0.8383	1912.72

Tabla 3. Resultados completos de las combinaciones de *frecuencia*, k y n mediante Grid Search.

El análisis de estos resultados permitió identificar la configuración óptima que maximiza la similitud de los datos sintéticos con los reales, es decir, donde la accuracy es la más baja. La combinación óptima se alcanzó con un *frecuencia* de 5×10^{-5}, $k = 200$, y $n = 50$, obteniendo una exactitud de 0.8200. Esta baja exactitud indica que el modelo tuvo dificultades

para distinguir entre los datos reales y los sintéticos, lo cual sugiere que los datos generados eran altamente realistas.

Además, se observa que esta combinación óptima tuvo el segundo mayor tiempo de ejecución (1841.10 segundos). Aunque el incremento en el tiempo de ejecución es considerable en comparación con configuraciones con valores menores de k y n, este aumento es aceptable dado el beneficio en la calidad de los datos sintéticos generados.

Los resultados completos permiten evaluar cómo los distintos valores de k y n afectan el rendimiento del modelo. Se observa que incrementando el número de *clusters* (k) y seleccionando un número moderado de diagnósticos ($n = 50$), se generan datos sintéticos más realistas. Valores altos de k permiten capturar mejor la heterogeneidad de los datos, mientras que un n moderado evita el ruido excesivo de incluir demasiados diagnósticos poco frecuentes.

A medida que se incrementa el número de *clusters* (k), se observa que el tiempo de ejecución aumenta considerablemente, como se ilustra en la Figura 8 (ver Anexo A). Aunque el tiempo de computación sigue una tendencia ascendente con respecto a k, este incremento no es cúbico, como se sugería en estudios previos como el de Aviñó *et al* [1]. Esto se debe a que el tiempo de ejecución refleja no solo el proceso de clustering, sino también la extracción de métricas y la creación de los datos sintéticos. Debido a la inclusión de estos pasos adicionales, el comportamiento del tiempo de ejecución difiere del esperado en un análisis que contemplara exclusivamente el proceso de clustering. Aun así, el algoritmo resulta relativamente eficiente en comparación con lo esperado en escenarios más complejos o con mayor cantidad de datos.

6.4 Comparación de las Variables más Importantes: Sintéticos vs. Reales

Para evaluar la calidad de los datos sintéticos en comparación con los reales, se realizaron pruebas siguiendo el enfoque del artículo de Aviñó *et al* [1]. Se utilizó un modelo de Random Forest para analizar las siguientes categorías de variables:

- **Atributos demográficos:** Información básica como edad y sexo.
- **Circunstancias de ingreso y alta:** Características como el tipo de ingreso y la disposición al alta.
- **Códigos diagnósticos ICD-9 y tratamientos:** La presencia de diagnósticos y tratamientos específicos para cada paciente.

Finalmente, se llevaron a cabo comparaciones detalladas de las variables más importantes entre los datos sintéticos y los reales. Mediante gráficos de densidad para variables continuas y gráficos de barras para variables categóricas, se evaluó la fidelidad de los datos sintéticos al replicar las distribuciones observadas en los datos reales. Estas comparaciones aseguran que los datos generados conservan las características clave de los datos reales, manteniendo la relevancia de las variables importantes tanto en los datos reales como en los sintéticos.

En la Figura 6, se muestra el impacto relativo de las 20 características más importantes que influyeron en el modelo de Random Forest para distinguir entre los datos reales y los sinté-

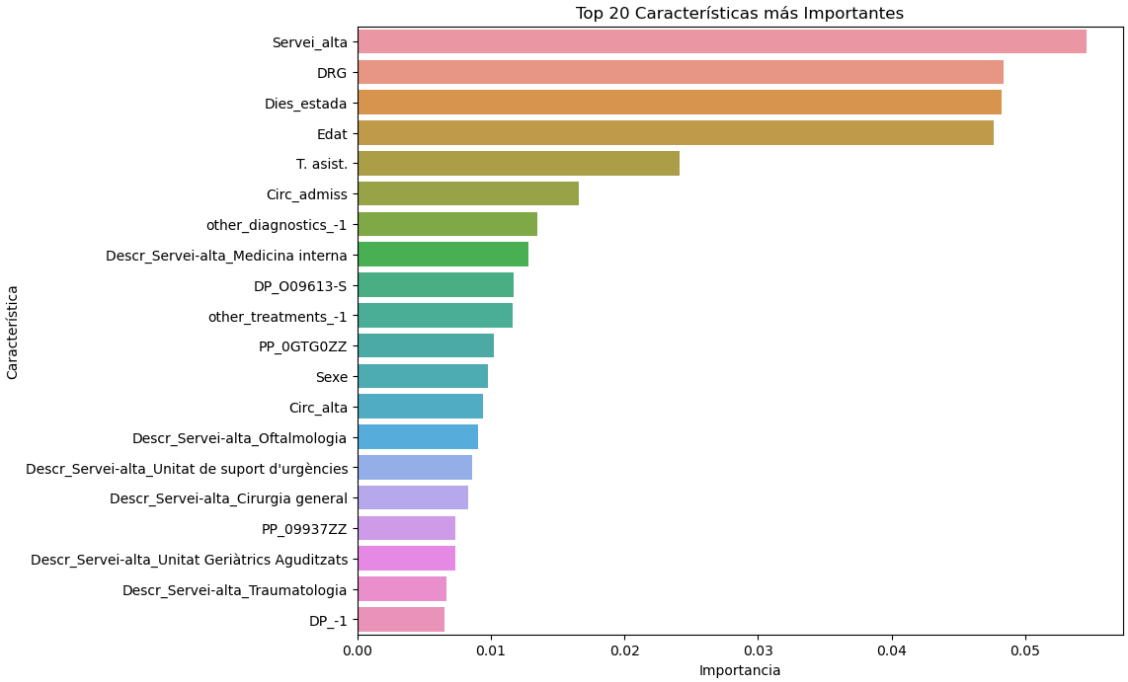

Figura 6. Top 20 características más importantes en el modelo de Random Forest.

ticos. Las variables más relevantes incluyen *Servei_alta*, *DRG*, *Dies_estada* y *Edat*, lo que indica que estas características jugaron un papel crucial en la clasificación de los episodios. La alta importancia de estas variables destaca su relevancia en los patrones de los datos, y su correcta replicación en los datos sintéticos asegura una generación de datos coherente.

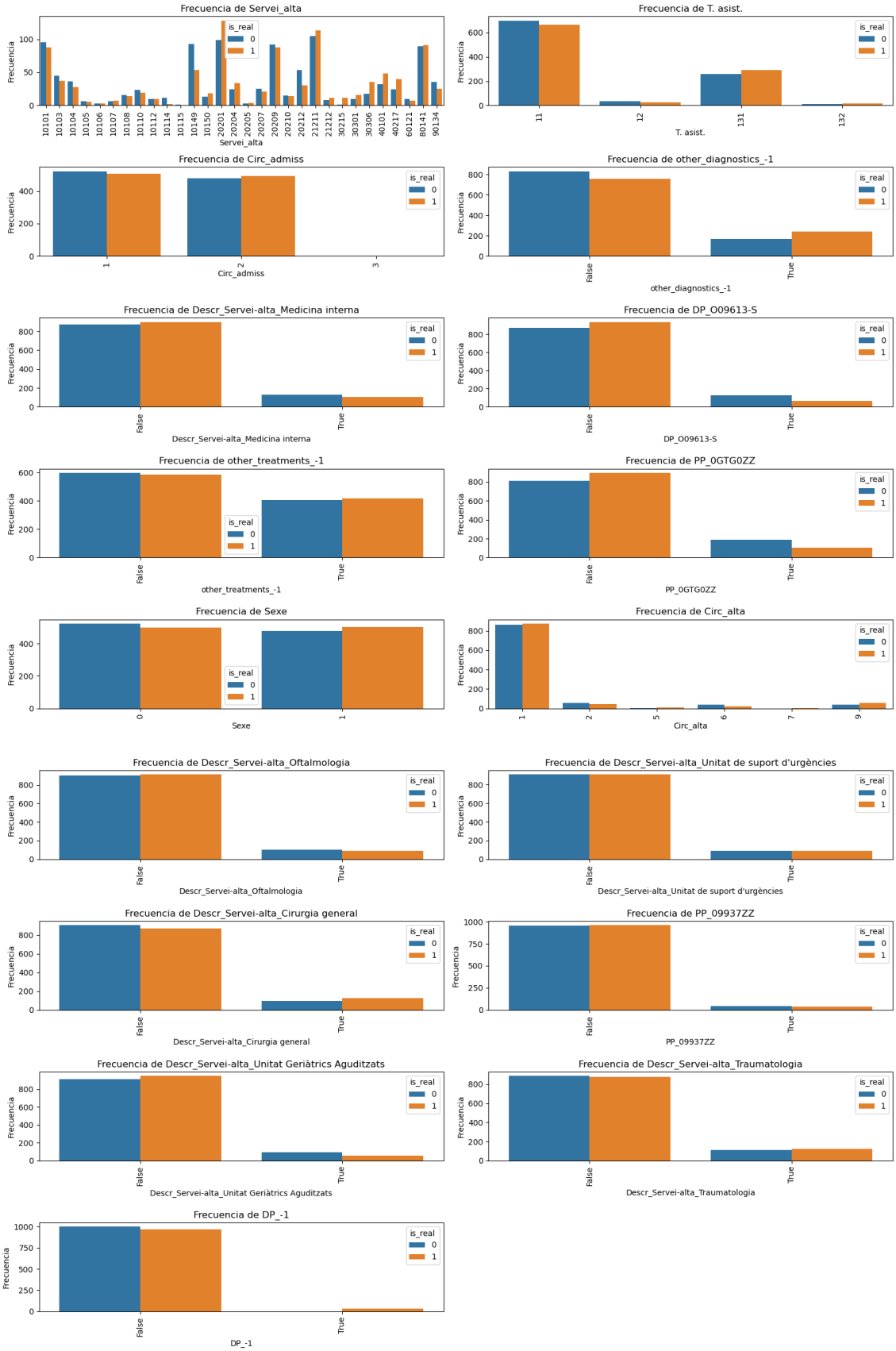

Figura 7. Comparación de las distribuciones de variables categóricas entre datos sintéticos y reales.

En la Figura 7, se realiza una comparación directa de las distribuciones de frecuencia de las 20 características categóricas más importantes entre los datos reales y los sintéticos. La mayoría de las variables muestran distribuciones muy similares, lo que refleja que los datos sintéticos replican de manera efectiva los patrones observados en los datos reales. Sin embargo, en algunos casos, como *Servei_alta*, se observan ligeras diferencias en las frecuencias de ciertas categorías, lo que sugiere que estas áreas podrían ser mejoradas para aumentar la exactitud de la generación de datos.

En conjunto, estos resultados confirman que los datos sintéticos generados son en su mayoría fieles a las características de los datos reales, lo que valida el enfoque de generación de datos sintéticos empleado. No obstante, las pequeñas discrepancias observadas en algunas variables categóricas sugieren posibles oportunidades de refinamiento en futuros experimentos para mejorar aún más la fidelidad de los datos generados.

Además de las pruebas estándar, se realizaron análisis adicionales introduciendo los diagnósticos *POA* (*Present On Admission*) en el modelo de Random Forest. Este enfoque mejoró considerablemente la precisión del modelo, facilitando la distinción entre los datos reales y los sintéticos con una accuracy de 0.93. Como se observa en la Figura 9 (ver Anexo B), las 20 características más relevantes estaban dominadas por los códigos *POA*, lo que sugiere que el orden y la presencia de estos diagnósticos es un factor clave que diferencia los dos conjuntos de datos. Este hallazgo subraya la importancia de modelar correctamente no solo las características de los datos, sino también su secuencia y orden de aparición, especialmente en entornos clínicos donde la cronología de los eventos médicos juega un rol crucial.

En la Figura 10 (ver Anexo B), se comparan las distribuciones de las 20 características categóricas más importantes entre los datos reales y los sintéticos. A diferencia de los resultados previos, aquí se observan algunas diferencias más marcadas, lo que sugiere que las variables relacionadas con los diagnósticos *POA* son particularmente difíciles de replicar con precisión. Estos resultados indican que la presencia y el orden de los diagnósticos al momento de la admisión del paciente son determinantes importantes para la fidelidad de los datos generados, y representan un área clave para futuros esfuerzos de mejora en la generación de datos sintéticos.

Para finalizar, y con el objetivo de comprobar que el método de clusterización y generación de datos replicaba los resultados de Aviñó *et al.*, generando datos de calidad para los pacientes, realizamos una validación utilizando Random Forest únicamente con los datos demográficos y los diagnósticos. Obtuvimos buenos resultados con una precisión (accuracy) de 0.56 (ver Anexo C), lo que confirma la efectividad del enfoque empleado.

Capítulo 7. Conclusiones

7.1 Conclusiones del Proyecto

El presente proyecto se centró en el desarrollo de un algoritmo para la generación de datos sintéticos que replicase las características de los datos reales del *CMBD* (Conjunto Mínimo Básico de Datos) para hospitales de agudos en Cataluña. El objetivo principal fue garantizar que los datos sintéticos mantuvieran la utilidad y las propiedades estadísticas de los datos reales, a la vez que se asegurara la privacidad de los pacientes.

Objetivos Específicos y Resultados Finales

Los resultados obtenidos cumplen con los objetivos planteados, destacando los siguientes puntos clave:

- **Calidad de los Datos Sintéticos:** El conjunto de datos sintéticos generado refleja adecuadamente las características esenciales y las distribuciones estadísticas de los datos reales. Este conjunto de datos mantiene la coherencia en las distribuciones de variables categóricas y continuas, siendo especialmente relevante para su uso en estudios futuros sin comprometer la privacidad de los pacientes.
- **Replicación de Resultados Previos:** Para confirmar que nuestro método replicaba los resultados de Aviñó *et al.* y generaba datos de calidad para pacientes, realizamos una validación con Random Forest utilizando solo datos demográficos y diagnósticos. Obtuvimos una precisión de 56 % (ver Anexo C), lo que confirma la efectividad del enfoque y su capacidad para capturar las características esenciales de los datos reales en pacientes, aunque no tanto en episodios de estos.
  Sin embargo, es importante destacar que el procedimiento funciona bien para pacientes pero no para episodios de estos. Esto sugiere que, aunque el modelo es efectivo en replicar las características generales de los pacientes, existen desafíos adicionales al intentar replicar con precisión los detalles de cada episodio clínico. Esta diferencia entre pacientes y episodios representa un área clave para futuras mejoras en el modelo de generación de datos sintéticos.
- **Evaluación mediante el Algoritmo Random Forest para Diferenciación de Datos:** Adicionalmente, el análisis mediante Random Forest mostró que el modelo alcanzó una exactitud del 82 % al distinguir entre los datos reales y los sintéticos cuando se consideraron todas las variables. Este resultado refleja que, si bien los datos sintéticos replican en buena medida las características de los reales, el algoritmo fue capaz de identificar diferencias entre ambos conjuntos. Aunque una exactitud cercana al 50 % sería ideal, ya que indicaría que los datos sintéticos son prácticamente indistinguibles de los reales, estos datos aún presentan un buen nivel de realismo, con margen de mejora en futuras iteraciones.
- **Comprobación de las Variables Más Importantes:** Se evaluaron las variables más influyentes en el modelo mediante gráficos de densidad y de barras, lo que permitió verificar que variables clave como *Servei_alta*, *DRG*, *Dies_estada* y *Edat* mostraron

distribuciones similares entre los datos reales y los sintéticos. Esto refuerza la validez del conjunto de datos sintéticos, aunque se observa que algunas diferencias entre los conjuntos podrían ser optimizadas.

- **Importancia de los Diagnósticos POA:** Un hallazgo significativo fue la dificultad para replicar con precisión las variables relacionadas con los diagnósticos *POA* (*Present On Admission*). Al incorporar estas variables en el modelo, se observó que mejoraba considerablemente su capacidad para distinguir entre los datos reales y los sintéticos llegando a un 93 % (Ver Anexo B). Esto indica que la presencia y el orden de los diagnósticos al momento de la admisión son factores determinantes en la fidelidad de los datos generados, representando un área clave para futuras mejoras.

Este proyecto ha logrado desarrollar un algoritmo eficaz para la generación de datos sintéticos que, en gran medida, replican fielmente las características y distribuciones observadas en los datos reales del *CMBD*. A pesar de que la exactitud del Random Forest fue superior al 50 %, lo que indica que aún se pueden mejorar algunos aspectos del conjunto sintético, los resultados obtenidos son positivos y útiles para futuras investigaciones.

A pesar de que existen diferencias entre los conjuntos de datos reales y sintéticos, los datos generados presentan una utilidad considerable en pruebas preliminares y simulaciones, sin comprometer la privacidad de los pacientes. El éxito en mantener la coherencia en las distribuciones individuales y en la estructura general de los *clusters* es un logro significativo que respalda el uso de estos datos en investigaciones futuras y el desarrollo de modelos predictivos.

7.2 Conclusiones Personales

El desarrollo de este proyecto ha sido una experiencia enriquecedora, permitiéndome explorar el campo de la inteligencia artificial aplicada al sector de la salud. A través del desarrollo de datos sintéticos, he aprendido no solo sobre el manejo ético de la información médica, sino también sobre los desafíos técnicos de replicar las relaciones entre múltiples variables clínicas en un entorno de datos sintéticos.

Uno de los mayores desafíos ha sido garantizar que los datos sintéticos preserven tanto las distribuciones univariantes como las relaciones interdependientes entre las variables clínicas y los episodios de cada paciente. Capturar estas conexiones es crucial para asegurar que los datos sintéticos reflejen no solo la variabilidad individual, sino también el curso clínico completo de los pacientes, algo que se puede mejorar en trabajos futuros.

Trabajar con datos reales del sector de la salud me ha proporcionado una visión más clara de cómo la tecnología puede transformar los sistemas de salud, siempre respetando la privacidad de los pacientes. Este proyecto ha consolidado mis conocimientos técnicos y ha reforzado mi compromiso con la investigación en tecnologías que pueden tener un impacto tangible en la sociedad.

Capítulo 8. Discusión y Trabajo Futuro

8.1 Limitaciones

El desarrollo de este proyecto reveló varias limitaciones, particularmente derivadas de la complejidad de los datos clínicos y las relaciones entre ellos. Mi supervisor, R. Gavaldà, me propuso seguir la metodología del trabajo de Aviñó *et al.* [1], que trabajaba con pacientes y episodios de manera conjunta, para investigar si la clusterización de pacientes permitiría generar episodios realistas. En este proyecto, la clusterización se realizó a nivel de pacientes, pero la generación de datos fue a nivel de episodios. Tras investigar esta aproximación, encontramos que la clusterización a nivel de pacientes genera buena información propia del paciente (ver Anexo C); sin embargo, si queremos generar episodios con información detallada y realista, deberíamos trabajar directamente a nivel de episodios. Esta decisión pudo haber afectado la capacidad del modelo para capturar la interrelación entre episodios de un mismo paciente. La representación secuencial de eventos médicos dentro de un mismo individuo se vio comprometida, especialmente en casos de pacientes que regresan al hospital por complicaciones o evolución de un diagnóstico previo. Esta dependencia entre episodios es especialmente relevante en enfermedades crónicas o tratamientos continuados, donde un episodio está directamente relacionado con el anterior. De esta forma, hemos dado respuesta a una de las hipótesis propuestas por nuestro supervisor, orientando las futuras líneas de trabajo hacia un enfoque centrado en episodios.

Uno de los principales desafíos fue representar la interrelación entre episodios dentro de un mismo paciente. Aunque se lograron reproducir con precisión las distribuciones univariantes y se consiguió una clusterización efectiva, el modelo no fue capaz de capturar las relaciones complejas entre múltiples episodios. Un ejemplo claro es la dificultad para emular correctamente la secuencia de visitas de un paciente con una afección crónica, donde un diagnóstico inicial tiene un impacto directo en episodios futuros.

Además, los resultados adicionales indicaron que al considerar el indicador *POA* (*Present On Admission*) en el modelo, la capacidad del Random Forest para distinguir entre datos reales y sintéticos aumentaba considerablemente. Aunque el estándar CMBD no especifica un orden particular en la entrada de diagnósticos secundarios, desconocíamos si en la práctica los hospitales los ingresan siguiendo algún orden especial. Al analizar los *POAs*, observamos que efectivamente existe un orden implícito, lo que sugiere que la secuencialidad y el orden en que aparecen los diagnósticos son factores críticos para una mejor replicación de la realidad clínica. Esta observación responde a otra de las preguntas planteadas por nuestro supervisor y marca una dirección clara para futuras líneas de investigación.

Otra limitación significativa fue el tiempo computacional requerido para la ejecución completa del algoritmo, que incluía no solo la clusterización, sino también la extracción de métricas y la generación de datos sintéticos. A medida que se incrementaba el valor de k, la complejidad y el tiempo de ejecución del modelo aumentaban considerablemente. Este incremento no fue cúbico tal y como se describe en el trabajo previo, debido a que, además de la clusterización, los cálculos adicionales para extraer métricas de cada *cluster* y la creación de los datos sinté-

ticos influían en el tiempo de computación. Durante la fase de búsqueda de hiperparámetros (*Grid Search*), este aumento en el tiempo de ejecución fue aún más pronunciado, ya que cada iteración requería la evaluación completa del modelo, lo que se tradujo en tiempos de ejecución elevados, particularmente para valores altos de k. Esta dificultad fue un obstáculo importante para explorar configuraciones de clusterización más complejas y optimizadas.

8.2 Futuras Líneas de Trabajo

Para abordar las limitaciones identificadas y mejorar la calidad y eficiencia del modelo, se proponen las siguientes líneas de trabajo futuro:

- **Modelar el Orden y Secuencialidad de los Diagnósticos y Tratamientos:** Una de las prioridades futuras será integrar el orden y la relación entre los diferentes diagnósticos y tratamientos en la generación de datos sintéticos. Esta mejora permitirá que los episodios consecutivos de un paciente reflejen mejor su trayectoria clínica, capturando la cronología de las visitas y los eventos médicos. Sin embargo, un desafío importante será evitar los problemas de sobredimensionalidad, ya que modelar adecuadamente la secuencialidad de diagnósticos y tratamientos puede generar un gran aumento en la cantidad de variables. Se deberá considerar la implementación de técnicas avanzadas de reducción de dimensionalidad o métodos de selección de características que preserven la información más relevante sin aumentar excesivamente la complejidad del modelo.
- **Clusterización por Episodios en Lugar de Pacientes:** Otro enfoque importante será trabajar con una clusterización basada en episodios y no únicamente en pacientes. Esto permitirá generar conjuntos de datos sintéticos que representen con mayor precisión la diversidad de situaciones clínicas que pueden ocurrir a lo largo de diferentes episodios de un paciente, mejorando la replicación de la continuidad del cuidado y su impacto en los resultados médicos.
- **Optimización de la Clusterización y Reducción del Tiempo de Cómputo:** El proceso de clusterización y extracción de métricas debe optimizarse para reducir los tiempos de ejecución, especialmente cuando se utilizan valores altos de k. Se investigarán técnicas más eficientes, como el uso de algoritmos de clusterización más escalables o la paralelización, para hacer el proceso más ágil y aplicable a conjuntos de datos más grandes y complejos sin comprometer la precisión de los resultados. Asimismo, se deberá optimizar la generación de datos sintéticos para que no incremente el tiempo de ejecución de manera tan drástica.
- **Incorporar Nuevos Indicadores y *POA* en el Modelo:** Los resultados preliminares indicaron que el uso de indicadores como el *POA* (*Present On Admission*) mejora la capacidad del modelo para distinguir entre datos reales y sintéticos. Por lo tanto, es necesario continuar explorando el impacto de estos indicadores en la calidad de los datos sintéticos y su utilidad en simulaciones clínicas junto con el estudio de la secuencialidad de los diagnósticos.

Estas cuatro líneas de trabajo son clave para superar las limitaciones actuales y mejorar la capacidad del modelo para generar datos sintéticos más realistas y aplicables. La implementación de estas propuestas permitirá avanzar hacia la creación de un sistema más robusto y

eficaz para la simulación y análisis de datos de episodios clínicos sin comprometer la privacidad de los pacientes.

Bibliografía

[1] Laura Aviñó, Matteo Ruffini, and Ricard Gavaldà. Generating synthetic but plausible healthcare record datasets. *arXiv preprint arXiv:1807.01514*, 2018. URL `https://arxiv.org/abs/1807.01514`.

[2] Matteo Ruffini, Ricard Gavaldà, and Esther Limón. Clustering patients with tensor decomposition. In *Proceedings of the Machine Learning for Healthcare (MLHC 2017)*. MLHC, August 2017.

[3] Giulio Ruffini, Stefanie Enriquez-Geppert, Kornelia D Kandylaki, Klaus Heilman, and Sonja Kuckertz. Tucker decomposition for tensor-based clustering of human brain patterns. In *2017 IEEE 27th International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6. IEEE, 2017.

[4] Wikipedia contributors. Random forest, 2024. URL `https://en.wikipedia.org/wiki/Random_forest`. Accessed: 2024-09-30.

[5] Richard J Chen, Ming Y Lu, Tiffany Y Chen, Drew F Williamson, and Faisal Mahmood. Synthetic data in machine learning for medicine and healthcare. *Nature Biomedical Engineering*, 5(6):493–497, 2021.

[6] Nicola Rieke, Jonny Hancox, Wenqi Li, Fausto Milletari, Holger R Roth, Shadi Albarqouni, Spyridon Bakas, Mathieu N Galtier, Bennett Landman, Klaus Maier-Hein, et al. The future of digital health with federated learning. *NPJ Digital Medicine*, 3(1):1–7, 2020.

[7] Edward Choi, Siddharth Biswal, Bradley Malin, Jon Duke, Walter F. Stewart, and Jimeng Sun. Generating multi-label discrete patient records using generative adversarial networks. In *Proceedings of the Machine Learning for Healthcare Conference*, pages 286–305. PMLR, 2017.

[8] Mohammad Kamrul Hasan Baowaly, Chih-Lin Lin, Chia-Hung Liu, and Kuan-Hao Chen. Synthesizing electronic health records using improved generative adversarial networks. *Journal of the American Medical Informatics Association*, 26(3):228–241, 2019.

[9] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems 27 (NIPS 2014)*, pages 2672–2680, 2014.

[10] Sanjeev Arora, Rong Ge, Yingyu Liang, Tengyu Ma, and Yi Zhang. Generalization and equilibrium in generative adversarial nets (gans). *arXiv preprint arXiv:1703.00573*, 2017. URL `https://arxiv.org/abs/1703.00573`.

[11] Jinsung Yoon, James Jordon, and Mihaela van der Schaar. Ganite: Estimation of individualized treatment effects using generative adversarial nets. In *International Conference on Learning Representations (ICLR)*, Vancouver, BC, Canada, 2018.

[12] Lei Xu, Maria Skoularidou, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni. Mode-

ling tabular data using conditional gan. In *Advances in Neural Information Processing Systems*, volume 32, pages 7335–7345, 2019.

[13] Stavroula Bourou, Andreas El Saer, Terpsichori Helen Velivassaki, Artemis Voulkidis, and Theodore Zahariadis. A review of tabular data synthesis using gans on an ids dataset. *Information*, 12(9):375, 2021. doi: 10.3390/info12090375.

[14] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein gan. *arXiv preprint arXiv:1701.07875*, 2017. URL https://arxiv.org/abs/1701.07875.

[15] Micha Frid-Adar, Itay Diamant, Eli Klang, Meir Amitai, Jonathan Goldberger, and Hayit Greenspan. Gan-based synthetic medical image augmentation for increased cnn performance in liver lesion classification. *Neurocomputing*, 321:321–331, 2018.

[16] Giovanni Geraci, Enrico Sepe, and Paolo Fornaro. Inaccuracy of international classification of diseases (icd) codes in identifying patients with complications after carotid endarterectomy. *Journal of Vascular Surgery*, 25(3):501–506, 1997.

[17] Matteo Ruffini. *Learning latent variable models: efficient algorithms and applications*. PhD thesis, Universitat Politècnica de Catalunya - BarcelonaTech, Barcelona, November 2019.

[18] Antonia Creswell, Tiffany White, Vincent Dumoulin, Kanishk Arulkumaran, Bharath Sengupta, and Anil Anthony Bharath. Generative adversarial networks: An overview. *IEEE Signal Processing Magazine*, 35(1):53–65, 2018.

[19] *Manual de Notificación de Hospitales Generales de Agudos*. Servei Català de la Salut (CatSalut), 2022. URL \url{https://catsalut.gencat.cat/web/.content/minisite/catsalut/proveidors_professionals/informacio_per_a_la_gestio/registre_activitat/cmbd/manuals/2022_manual_cmbd_ah.pdf}. Accessed: 2024-05-14.

Anexos

El código desarrollado en este proyecto estará disponible en un repositorio público de GitHub, donde podrá ser consultado, seguido y descargado por cualquier interesado. El código se distribuye bajo la licencia Creative Commons Attribution 4.0 International (CC BY 4.0), lo que permite su uso, distribución y modificación siempre que se otorgue el crédito adecuado a los autores.

Anexo A: Relación tiempo ejecución vs clusters

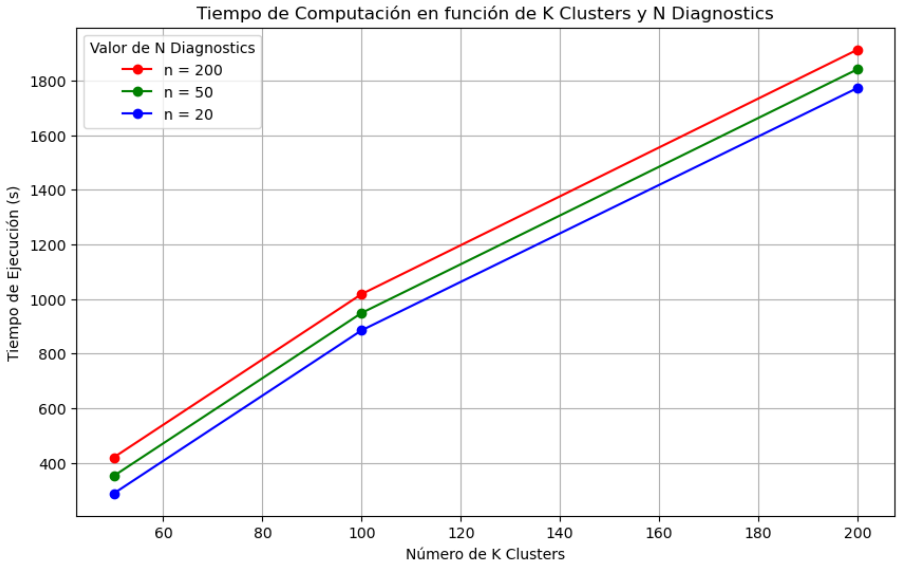

Figura 8. Relación entre el número de K Clusters y el tiempo de ejecución en segundos.

Este gráfico representará visualmente la correlación entre el número de clusters y el tiempo de ejecución, permitiendo una comparación clara de los resultados obtenidos en el proyecto con los reportados en estudios previos. Se observa una tendencia lineal en el aumento del tiempo conforme se incrementa el número de clusters.

Anexo B: Análisis Adicional de Diagnósticos *POA*

	Precision	Recall	F1-Score	Accuracy
Results with POA	0.93	0.93	0.93	0.93

Tabla 4. Resultados Generales del Informe de Clasificación con POA

En esta tabla (Tabla 4) se presentan los resultados detallados del análisis con las variables *POA*, que tuvieron una relevancia destacada en la diferenciación entre datos reales y sintéticos. A continuación se muestran las 20 características más importantes (Figura 9) y la

comparación de sus frecuencias entre los conjuntos de datos reales y sintéticos (Figura 10).

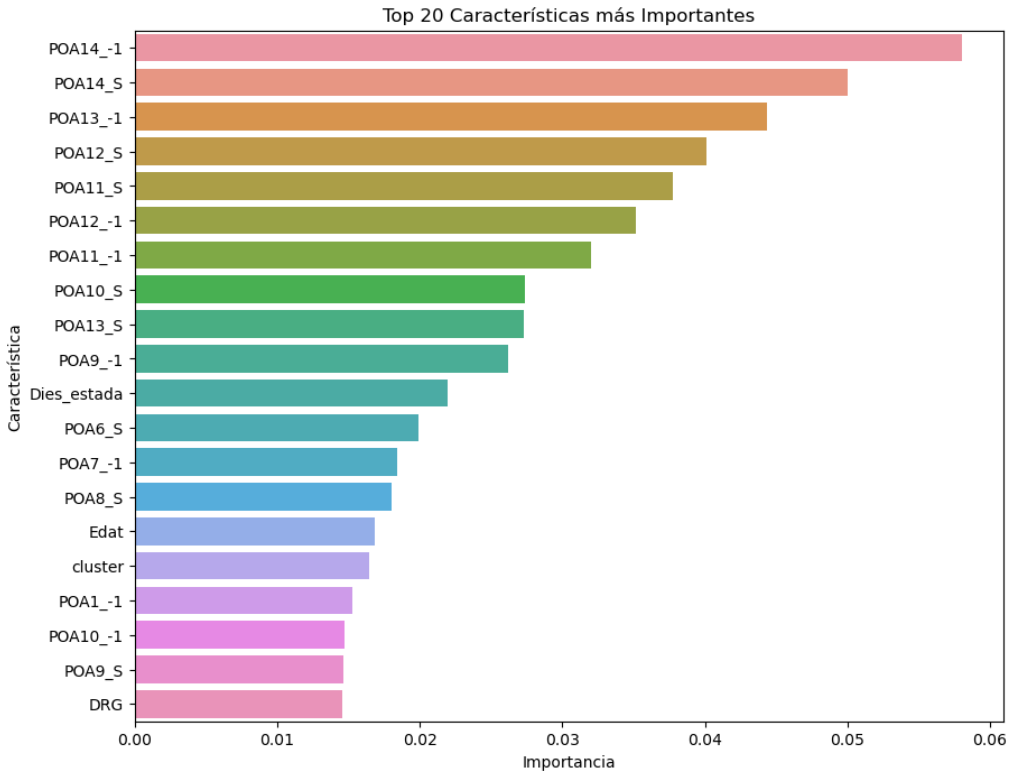

Figura 9. Top 20 características más importantes en el análisis de *Random Forest* con diagnósticos *POA*.

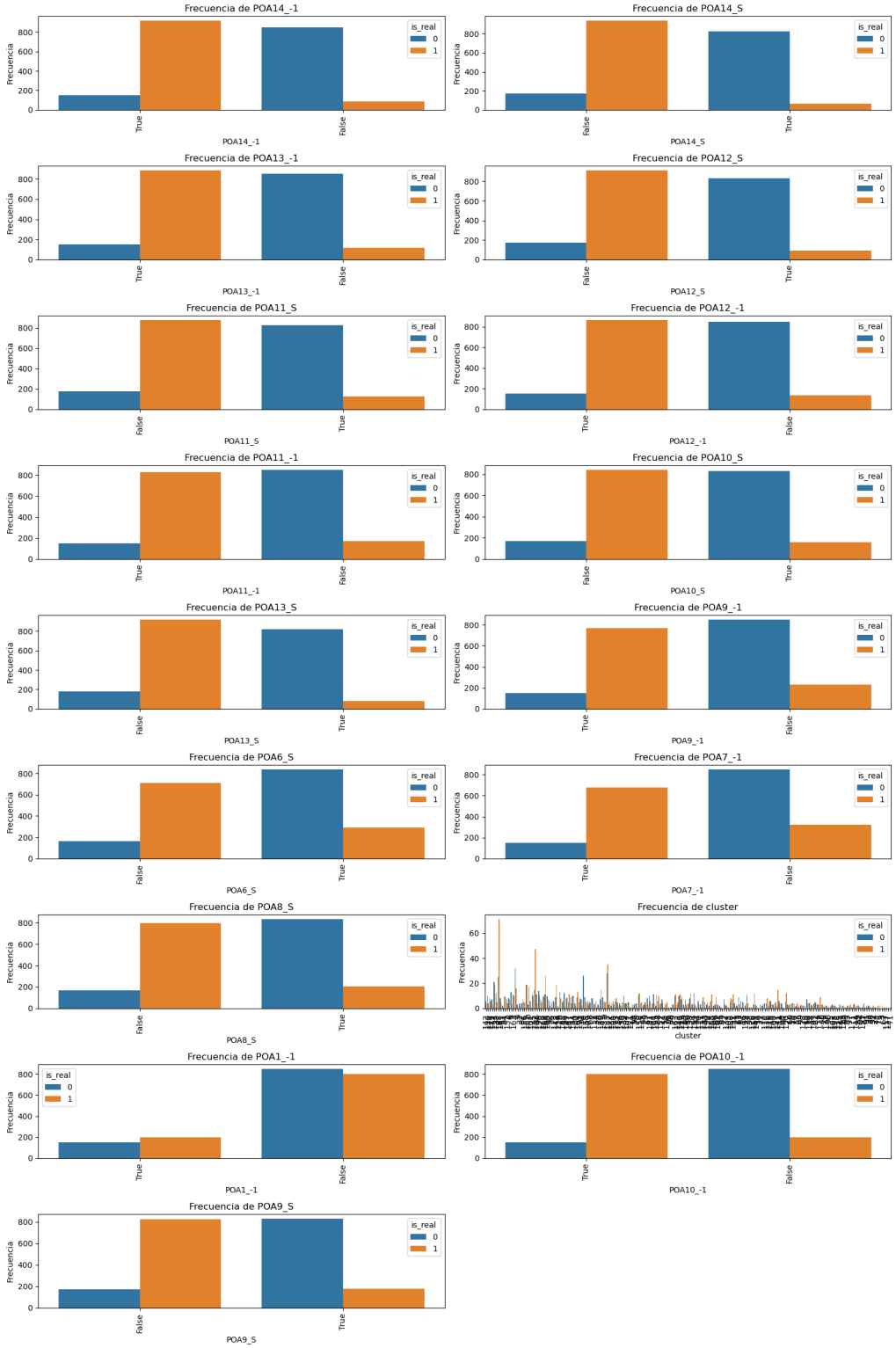

Figura 10. Comparación de las frecuencias de las 20 características categóricas más importantes entre los datos reales y sintéticos con diagnósticos *POA*.

Se observan algunas diferencias significativas, lo que sugiere la dificultad de replicar con precisión las características relacionadas con el *POA*.

Anexo C: Análisis Adicional de Variables Demográficas y Diagnósticos

	Precision	Recall	F1-Score	Accuracy
Demographics and Diagnostics	0.56	0.56	0.55	0.56

Tabla 5. Resultados Generales del Informe de Clasificación con Variables Demográficas y Diagnósticos

La Tabla 5, muestra los resultados generales del análisis realizado con variables demográficas y diagnósticos. La precisión (accuracy) de 0.56 indica una baja capacidad para distinguir entre datos reales y sintéticos, lo cual refleja un buen desempeño de acuerdo con la metodología empleada.

A continuación, en la Figura 11, se muestran las 20 características más importantes identificadas por el modelo Random Forest. Como se puede observar, la variable Edad destaca con una gran relevancia, seguida de otras características tanto demográficas como diagnósticos. La importancia relativa de cada característica se indica con barras de distinto tamaño, lo que permite identificar las variables clave para el modelo.

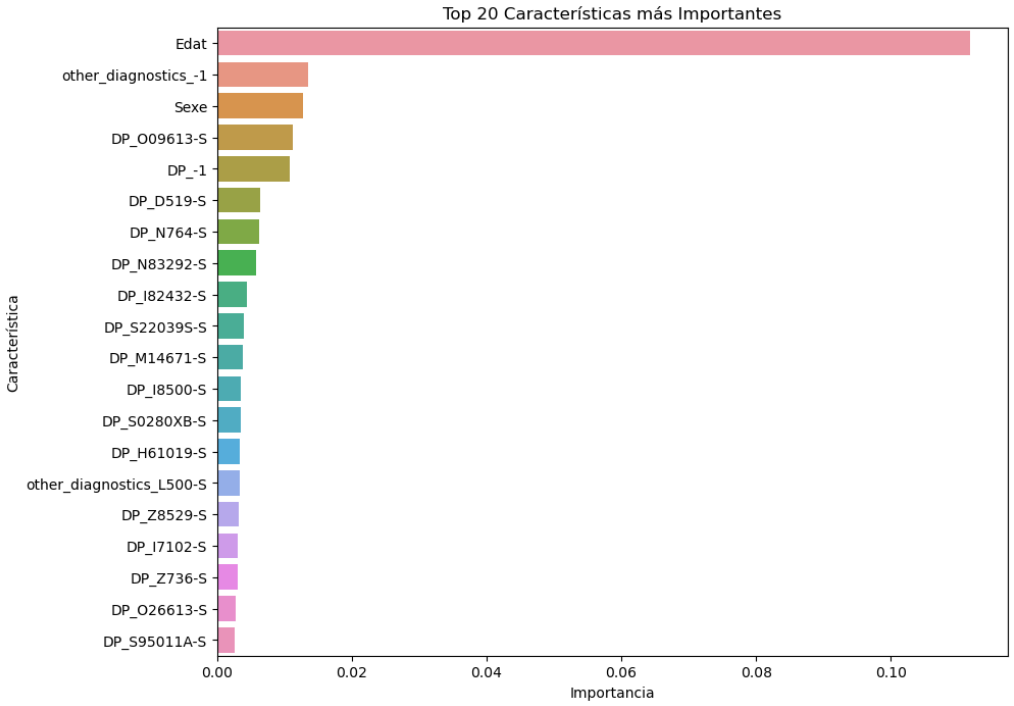

Figura 11. Top 20 características más importantes en el análisis de *Random Forest* con variables demográficas y diagnósticos.

En la Figura 12, se presenta una comparación de las frecuencias de estas 20 características categóricas entre los datos reales y sintéticos. Aunque la mayoría de las variables muestran distribuciones similares, algunas diferencias pueden observarse, lo que indica áreas potenciales de mejora en la generación de datos sintéticos.

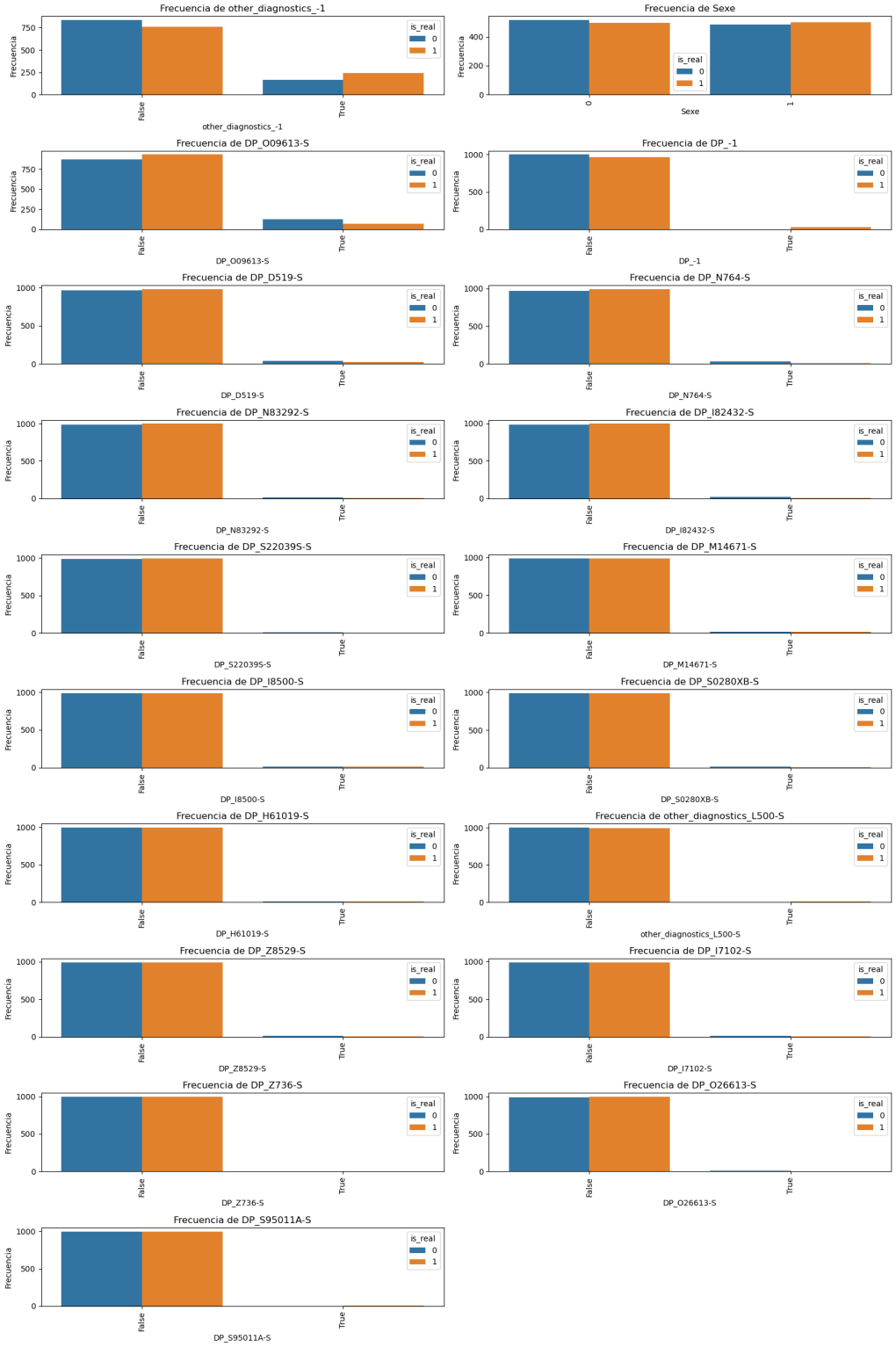

Figura 12. Comparación de las frecuencias de las 20 características categóricas más importantes entre los datos reales y sintéticos con variables demográficas y diagnósticos.