UNIVERSITAT POLITÈCNICA DE VALÈNCIA

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Escuela Técnica Superior de Ingeniería Industrial

Diseño e implementación de un proceso de desarrollo de IA confiable en salud

Trabajo Fin de Grado

Grado en Ingeniería Biomédica

AUTOR/A: Manuel Vicente, Carlos de

Tutor/a: Sáez Silvestre, Carlos

CURSO ACADÉMICO: 2023/2024

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

ESCOLA TÈCNICA SUPERIOR ENGINYERIA INDUSTRIAL VALÈNCIA

TRABAJO FIN DE GRADO EN INGENIERÍA BIOMÉDICA

DISEÑO E IMPLEMENTACIÓN DE UN PROCESO DE DESARROLLO DE IA CONFIABLE EN SALUD

AUTOR: CARLOS DE MANUEL VICENTE

TUTOR: CARLOS SÁEZ SILVESTRE

Curso Académico: 2023-2024

Agradecimientos

A todas las personas que han hecho posible este trabajo, mi más sincero agradecimiento. Agradezco al Ministerio de Educación por la Beca de colaboración que ha permitido realizar esta investigación. A Carlos, por su invaluable guía, apoyo y confianza a lo largo de este proyecto. Su experiencia, consejos y atención han sido fundamentales para el desarrollo de mi investigación. Gracias por darme la oportunidad de participar en el laboratorio.

A mis compañeros de etapa, Nacho, Cris, Vicent, Marta, Joel y Luan, por su constante apoyo, amistad y por compartir conmigo este camino. Sin vosotros, esta experiencia no habría sido la misma. Gracias por estar siempre ahí.

A mi familia, por su amor incondicional y su constante respaldo. Este trabajo cierra una etapa significativa en mi vida e intenta proyectar la educación en valores y excelencia que me habéis transmitido. Gracias por estar siempre a mi lado, por enseñarme a perseverar y a creer en mis capacidades.

Resumen

Las tendencias tecnológicas señalan a la Inteligencia Artificial (IA) como una herramienta crucial en nuestra sociedad, pero su desarrollo debe respetar los derechos humanos. A pesar de la disponibilidad de información al respecto, los desarrolladores carecen de una guía práctica para abordar la construcción de IA confiable. Por tanto, este proyecto tiene como objetivo principal desarrollar un guía o pipeline que sirva como estructura para la creación de sistemas de IA. Los métodos se implementarán utilizando lenguajes de programación de vanguardia como Python y se validarán con conjuntos de datos representativos de la realidad. Se espera que este trabajo proporcione a cualquier desarrollador de IA una estructura teórica y ejemplos prácticos de los requisitos necesarios para crear sistemas confiables. Los resultados de esta investigación ofrecen una metodología gráfica para la construcción de una IA confiable mediante el uso de una matriz que expresa los métodos técnicos a emplear para cada requisito a lo largo del ciclo de vida de la IA. Asimismo, se ofrece una demostración práctica mediante la ejemplificación con código de la validación del pipeline con conjuntos de datos en salud.

Carlos de Manuel Vicente

cdeman@etsii.upv.es

Palabras clave: inteligencia artificial confiable, software, aprendizaje automático.

Resuma

Les tendències tecnològiques assenyalen a la Intel·ligència Artificial (IA) com una ferramenta crucial en la nostra societat, però el seu desenvolupament ha de respectar els drets humans. Malgrat la disponibilitat d'informació sobre aquest tema, els desenvolupadors manquen d'una guia pràctica per a abordar la construcció de IA de confiança. Per tant, aquest projecte té com a objectiu principal desenvolupar un guia o pipeline que servisca com a estructura per a la creació de sistemes de IA. Els mètodes s'implementaran utilitzant llenguatges de programació d'avantguarda com Python i es validaran amb conjunts de dades representatives de la realitat. S'espera que aquest treball proporcione a qualsevol desenvolupador de IA una estructura teòrica i exemples pràctics dels requisits necessaris per a crear sistemes de confiança. Els resultats d'esta investigació oferixen una metodologia gràfica per a la construcció d'una IA de confiança mitjançant l'ús d'una matriu que expressa els mètodes tècnics a emprar per a cada requisit al llarg del cicle de vida de la IA. Així mateix, s'oferix una demostració pràctica mitjançant l'exemplificació amb codi de la validació del pipeline amb conjunts de dades en salut.

Carlos de Manuel Vicente

cdeman@etsii.upv.es

Paraules clau: intel·ligència artificial de confiança, programari, aprenentatge automàtic.

Abstract

Technological trends point to Artificial Intelligence (AI) as a crucial tool in our society, but its development must respect human rights. Despite the availability of information in this regard, developers lack a practical guide to address the construction of reliable AI. Therefore, the main objective of this project is to develop a guideline or pipeline to serve as a structure for the creation of AI systems. The methods will be implemented using state-of-the-art programming languages such as Python and validated using representative real-world datasets. It is hoped that this work will provide any AI developer with a theoretical framework and practical examples of the requirements needed to create reliable systems. The results of this research provide a graphical methodology for building a reliable AI by using a matrix that expresses the technical methods to be employed for each requirement throughout the AI lifecycle. In addition, a practical demonstration is provided by exemplifying with code the validation of the pipeline with healthcare datasets.

Carlos de Manuel Vicente

cdeman@etsii.upv.es

Keywords: trustworthy artificial intelligence, pipeline, machine learning.

Índice General

DOCUMENTO I: MEMORIA

ÍNDICE DE FIGURAS **9**

ÍNDICE DE TABLAS **11**

ACRÓNIMOS **12**

CAPÍTULO 1. INTRODUCCIÓN **13**

1.1. MOTIVACIÓN 13
1.2. OBJETIVOS 16

CAPÍTULO 2. ANTECEDENTES **18**

2.1. MARCO EUROPEO DE LA IA CONFIABLE 18
2.2. REGULACIÓN DE LA IA 22
2.3. CICLO DE VIDA DE LA IA 23
2.4. APROXIMACIONES RELACIONADAS 27

CAPÍTULO 3. MATERIALES Y MÉTODOS **29**

3.1. METODOLOGÍA 29
3.2. HERRAMIENTAS 30
3.3. CONJUNTOS DE DATOS 31

CAPÍTULO 4. RESULTADOS **33**

4.1. MATRIZ DE IA CONFIABLE 33
4.2. PIPELINES SOFTWARE PARA IA CONFIABLE 35
4.3. EVALUACIÓN EN CONJUNTO DE DATOS “DIABETES” 43
4.4. EVALUACIÓN EN CONJUNTO DE DATOS “HEART DISEASE” 62
4.5. CHECKLIST DE RECOMENDACIONES PARA EL DESARROLLO DE IA CONFIABLE 72

CAPÍTULO 5. DISCUSIÓN **74**

5.1. IMPACTO 74
5.2. POSICIONAMIENTO EN EL ESTADO DEL ARTE 74
5.3. LIMITACIONES Y TRABAJO FUTURO 75

CAPITULO 6. CONCLUSIONES **77**

CAPÍTULO 7. BIBLIOGRAFÍA **78**

DOCUMENTO II: PRESUPUESTO

ÍNDICE DE TABLAS **85**

1. INTRODUCCIÓN **86**

2. CUADRO DE PRECIOS DE MANO DE OBRA **86**

3. CUADRO DE PRECIOS DE MAQUINARIA **86**

4. CUADRO DE PRECIOS UNITARIOS **87**

5. CUADRO DE PRECIOS DESCOMPUESTOS **88**

6. PRESUPUESTOS PARCIALES **89**

7. PRESUPUESTO TOTAL DE EJECUCIÓN POR CONTRATA **91**

DOCUMENTO I

Memoria

Índice de figuras

Figura 1. Marco de trabajo para alcanzar una IA confiable (European Comission, 2019). 16
Figura 2. Principales relaciones entre principios y requisitos de una IA confiable según la guía de la UE ... 18
Figura 3. Ciclo de vida de minería de datos según CRISP-DM (SPSS Modeler Subscription, 2021). 23
Figura 4. Ciclo de vida de una IA generativa (GenAI)(J. Saltz, 2024). 24
Figura 5. Ciclo de vida de un proyecto de machine learning en Amazon Web Services (AWS)(Well-Architected machine learning lifecycle - Machine Learning Lens, 2023). 25
Figura 6. Ciclo de vida de un sistema de aprendizaje máquina según MLOps (Neupane, 2023). 26
Figura 7. Flujograma de la metodología de trabajo. 29
Figura 8. Matriz de requisitos para una IA confiable según la etapa del ciclo de vida. 34
Figura 9. Gráfico de barras de valores perdidos con umbral de eliminación en 40% perdidos en el "Conjunto Diabetes". 45
Figura 10. Box-plot de las características numéricas normalizas antes y después de umbralizar en el "Conjunto Diabetes". 46
Figura 11. Box plot de características numéricas por clase. 46
Figura 12. Gráfico de barras de las frecuencias absolutas de cada categoría en la variable 'gender' en el "Conjunto Diabetes". 47
Figura 13. Gráfico de dispersión 3D de las tres primeras componentes principales en el "Conjunto Diabetes". 47
Figura 14. Proyección FAMD las 2 componentes principales de un subconjunto del "Conjunto Diabetes" 48
Figura 15. Distribución de datos por categoría y clase para cada variable sensible en el "Conjunto Diabetes". 49
Figura 16. Distribución de valores perdidos por categoría según las variables sensibles en el "Conjunto Diabetes". 49
Figura 17. Mapa de calor de la diferencia de las matrices de confusión obtenidas con las categorías 'Caucasian' y 'AfricanAmerican' en la variable sensible 'race' en el "Conjunto Diabetes". 50
Figura 18. Comparación curvas ROC para un modelo simple, con reponderación y con sobremuestreo en el "Conjunto Diabetes". 51
Figura 19. Predicciones de datos con distinta raza de modelo entrenado normal y con sobremuestreo.... 51
Figura 20. Importancia en la predicción de cada variable según la raza en el "Conjunto Diabetes". ... 52
Figura 21. Mapa de calor de la matriz de correlaciones de las variables numéricas en el "Conjunto Diabetes". 53
Figura 22. Porcentaje de muestras por clase y categoría en el "Conjunto Diabetes". 54
Figura 23. FAMD de 10000 puntos para todas las características del "Conjunto Diabetes" 54
Figura 24. Curvas ROC multiclase para modelos Random Forest y Naive Bayes en el "Conjunto Diabetes". 55
Figura 25. Matrices de confusión para los modelos Random Forest y Naive Bayes en el "Conjunto Diabetes". 55
Figura 26. Diagrama de barras apiladas con importancia de las variables según clase para el modelo Random Forest en el "Conjunto Diabetes". 56
Figura 27. Gráficos explicativos de la predicción para una clasificación binaria en el "Conjunto Diabetes". 56
Figura 28. Curvas ROC para diferentes modelos entrenados de forma óptima y balanceada en el "Conjunto Diabetes". 58

Figura 29. Matrices de confusión para varios modelos entrenados de forma robusta en el "Conjunto Diabetes" 58
Figura 30. Curvas de calibración y densidad de probabilidad para diferentes modelos en el “Conjunto Diabetes”: A) Modelo Random Forest, B) Modelo Naive Bayes..... 59
Figura 31. Distribución de probabilidad de los modelos Bootstrap para distintos datos del "Conjunto Diabetes". 60
Figura 32. Densidades de probabilidad de dos casos positivos perturbados del “Conjunto Diabetes”: A) Predicción segura, B) Predicción insegura..... 61
Figura 33. Distribuciones de probabilidad para datos anómalos y normales predichos con varios modelos bootstrap o perturbados del "Conjunto Diabetes”..... 62
Figura 34. Gráfico de barras de valores perdidos con umbral de eliminación en 40% perdidos en el “Conjunto Heart Disease”..... 64
Figura 35. Proyección de las 3 componentes principales tras realizar un PCA de las características numéricas del “Conjunto Heart Disease”. 64
Figura 36. Distribución de datos por categoría y clase para cada variable sensible en el “Conjunto Heart Disease”. 65
Figura 37. Distribución de los datos perdidos según el sexo 65
Figura 38. Mapa de calor de la diferencia de las matrices de confusión obtenidas con las categorías ‘male’ y ‘female’ en la variable sensible ‘sex’ en el “Conjunto Heart Disease”. 66
Figura 39. Comparación curvas ROC para un modelo simple, con reponderación y con sobremuestreo en el “Conjunto Heart Disease”. 66
Figura 40. Comparación curvas ROC de modelos entrenados con muestras de diferente sexo en el “Conjunto Heart Disease”..... 67
Figura 41. Importancia de las variables para modelos Random Forest entrenados con distintas categorías 67
Figura 42. Gráfico bivariante de las características numéricas en el “Conjunto Heart Disease”..... 68
Figura 43. Porcentaje de muestras por clase y categoría en el “Conjunto Heart Disease”. 69
Figura 44. Proyección 2D de FAMD para el "Conjunto Heart Disease" 69
Figura 45. Matrices de confusión para los modelos Random Forest y Naive Bayes en el “Conjunto Heart Disease”. 70
Figura 46. Comparación de las curvas ROC para Random Forest y Naive Bayes tras optimizar hiperparámetros y balancear las clases en el “Conjunto Heart Disease”..... 70
Figura 47. Curvas de calibración y densidad de probabilidad para los diferentes modelos en el “Conjunto Heart Disease”..... 71
Figura 48. Distribución de probabilidad de los modelos Bootstrap para distintos datos del "Conjunto Heart Disease". 71
Figura 49. Densidad de probabilidad de un caso positivo perturbado del “Conjunto Heart Disease” . 72

Índice de tablas

Tabla 1. Metadatos iniciales para el “conjunto Diabetes”. 44
Tabla 2. Ejemplo de registro de las métricas de un entrenamiento del modelo Random Forest para el “Conjunto Diabetes” 57
Tabla 3. Métricas conjuntas de 100 modelos Bootstrap para el "Conjunto Diabetes". 60
Tabla 4. Métricas para modelos entrenados con variabilidad temporal 62
Tabla 5. Metadatos iniciales para el “conjunto Heart Disease" 63
Tabla 6. Checklist de recomendaciones para una IA confiable según requisito y ciclo de vida (P: Data preparation, D: Model development, U: Deployment & Use, M: Management, N/A: Not applicable) 73

Acrónimos

AWS Amazon Web Services

CV Validación cruzada (del inglés Cross Validation)

CIE-9 Clasificación Internacional de Enfermedades 9ª Revisión

FAMD Análisis Factorial de Datos Mixtos (del inglés Factorial Analysis of Mixed Data)

HT2 Hotelling T^2

IA Inteligencia Artificial

KNN K Vecinos más próximos (del inglés K-Nearest Neighbor)

MDR Reglamento de Dispositivos Médicos (del inglés Medical Device Regulation)

MICE Imputaciones Múltiples mediante Ecuaciones Encadenadas (del inglés Multivariate Imputation by Chained Equations)

ML Aprendizaje Máquina (del inglés Machine Learning)

NB Naive Bayes

PCA Análisis de Componentes Principales (del inglés Principal Component Analysis)

RF Random Forest

ROC Característica Operativa del Receptor (del inglés Receiver Operating Characteristic).

SHAP Explicaciones Aditivas Shapley (del inglés SHapley Additive exPlanations)

UE Unión Europea

VSCode Visual Studio Code

CAPÍTULO 1. Introducción

1.1. Motivación

Durante los recientes años, la tecnología y, en especial la Inteligencia Artificial (IA), ha ganado relevancia debido a las posibilidades que brinda, especialmente en áreas tan cruciales como la salud. Los sistemas basados en IA son capaces de mejorar multitud de aspectos de la vida humana. Su versatilidad permite utilizar la IA en una amplia gama de aplicaciones en el ámbito de la salud, beneficiando tanto a profesionales como a pacientes al agilizar procesos y realizar predicciones precisas sobre diversas situaciones clínicas.

A pesar de encontrarse en pleno auge, los continuos avances en la IA se realizan en ocasiones a expensas de aspectos tan fundamentales como los derechos humanos. Particularmente, en el ámbito de la salud, la tolerancia al error es prácticamente nula, dado que la vida de las personas depende de la precisión y la rigurosidad de estos sistemas, y la presencia de información sensible es especialmente frecuente. Es, pues, en este contexto donde la IA Confiable (TAI, del inglés Trustworthy Artificial Intelligence) adquiere relevancia, asegurando que los sistemas de IA no solo sean efectivos, sino también seguros y éticos.

Así pues, el término de IA confiable hace referencia a aquel sistema de IA que, durante todo su ciclo de vida, se asegura de cumplir ciertos componentes y requisitos que respetan los derechos fundamentales. En el caso de la UE, estos requisitos se ajustan a la Carta de Derechos Fundamentales de la Unión Europea (Charter EU) (Charter of Fundamental Rights of the European Union, 2016; European Comission, 2019). En consecuencia, la Comisión Europea identifica tres componentes en la IA confiable (European Comission, 2019):

- **Legal:** los sistemas de IA operan en un mundo organizado en base a unas leyes que prohíben y habilitan acciones, por lo que debemos asegurarnos de que estos cumplen con la normativa vigente. En cierto modo, este componente es exclusivamente objetivo: tu sistema de IA debe de cumplir con el sistema legal.
- **Ética:** el ámbito legal puede no cubrir el ámbito ético, por ello es necesario asegurar que la IA esté alineada con los valores éticos subyacentes a los derechos fundamentales establecidos por la Carta de Derechos Fundamentales de la Unión Europea (Charter of Fundamental Rights of the European Union, 2016).
- **Robusta:** asegurar que el sistema no causa daño inintencionado. Emplear salvaguardas para asegurar un correcto rendimiento frente a factores inesperados, tanto desde una perspectiva técnica y social, es decir, considerando tanto la aplicación y el ciclo de vida como el contexto y entorno de trabajo.

Considerando pues un sistema de IA en salud, si atendemos a la Reglamento de Dispositivos Médicos (EU MDR, por sus siglas en inglés Medical Device Regulation) (Regulation (EU) 2017/745 on Medical Devices, 2017), que establece un marco regulatorio para la seguridad y rendimiento de los dispositivos médicos en la UE, y la Ley de la IA (EU AI Act) (Radley-Gardner et al., 2016), la cual clasifica los sistemas de salud como de Alto Riesgo, es pues indispensable atender a la confiabilidad para garantizar el cumplimiento de los estándares éticos y legales.

1.1.1. Derechos fundamentales como base de una IA confiable

Entre el conjunto indivisible de derechos establecidos por los Tratados y la Carta de la UE, la UE (European Comission, 2019) las siguientes familias como particularmente aptas para regular el funcionamiento y la correcta construcción de las IAs. Aunque algunos de estos derechos son jurídicamente exigibles y, por lo tanto, de cumplimiento obligatorio, también permiten realizar una reflexión ética para comprender cómo el ciclo de vida de la IA puede afectar a los derechos humanos.

- **Respeto de la dignidad humana:** todo ser humano tienen un valor intrínseco y todas las personas deben ser tratadas de forma respetuosa, como sujetos morales y no como objetos manipulables.
- **Libertad del individuo:** todo ser humano debe ser libre para tomar decisiones. Implica no solo la libertad de la intrusión soberana, sino también la intervención gubernamental y no gubernamental para asegurar acceso equitativo a las personas en riesgo de exclusión. La libertad del individuo requiere de la mitigación de todas las acciones directa o indirectamente ilegítimas: amenazas a la autonomía mental, salud mental, vigilancia injustificada, engaño y manipulación desleal.
- **Respecto a la democracia, justicia y el Estado de Derecho:** todo poder gubernamental debe estar legalmente autorizado y limitado por la ley. Las IAs deben respetar la pluralidad de valores y libertad individual, en ningún caso afectar a procesos democráticos, deliberación o humana o sistemas de voto.
- **Equidad, no discriminación y solidaridad** – ***incluyendo los derechos de personas en riesgo de exclusión*****:** Asegurar el mismo respeto del valor moral y la dignidad de todos los seres humanos, más allá de la no discriminación. En el contexto de la IA, la equidad atañe a la incapacidad de generar salidas segadas injustamente. También requiere respeto para las personas y grupos potencialmente vulnerables.
- **Derechos de los ciudadanos:** las IAs ofrecen el potencial de escalar y mejorar la eficiencia gubernamental en la provisión de los bienes y servicios de los cuales se benefician los ciudadanos. Del mismo modo, sus derechos pueden verse afectados negativamente. El término "derecho de los ciudadanos" no pretende desatender los derechos de nacionales de terceros países y personas irregulares, que también cuentan con derechos internacionales y deben considerarse en el ámbito de la IA

1.1.2. Principios éticos

La UE lista cuatro principios éticos para simplificar el cumplimiento con los derechos fundamentales. Se remarca el hecho de que no existe ningún tipo de prioridad ni relación jerárquica entre ellos, por lo que, a la hora de construir un sistema, se deben de buscar satisfacer con igual relevancia.

- **Respeto de la autonomía humana:** principio basado en el empoderamiento de las capacidades cognitivas, sociales y culturales de los humanos. El diseño debe ser centrado en el humano de modo que estos mantengan su autodeterminación y en ningún caso se vean subordinados o condicionados

- **Prevención de daños:** principio que busca proteger la integridad física y mental de los humanos. Se debe prestar atención a aquellas situaciones donde el daño puede aumentarse por asimetrías de poder o información. Este principio atañe tanto al entorno natural como a todos los seres vivos.
- **Justicia:** principio definido por dos dimensiones: sustantiva y procedural. La primera atañe a la distribución equitativa de beneficios y coste al tiempo que se asegure la inexistencia de sesgo, discriminación o estigmatización tanto individual como grupal. La procedural implica impugnar e identificar responsables en las decisiones de IAs o humanos.
- **Explicabilidad:** principio cuyo objetivo es procurar transparencia en los procesos de modo que las decisiones sean explicables. Establece una comunicación abierta de las capacidades y propósitos de la IA. El grado de explicabilidad se ajustará a la aplicación y la gravedad de una salida errónea.

1.1.3. Tensiones entre principios

Cabe destacar que, a priori, no existe ningún tipo combinación de métodos o sistema perfecto que supongan una solución ideal para optimizar al máximo todos los principios. Existen tensiones entre ellos, por lo que el aumento de uno puede significar la disminución de otro, como es el caso de la prevención de daños y el respeto de la autonomía humana. Por ejemplo, puede surgir una tensión cuando una IA sugiere un tratamiento más agresivo para prevenir posibles daños, mientras que el paciente prefiere un enfoque menos agresivo.

Así pues, para cada contexto y aplicación específica se deben analizar todas las situaciones, de modo que se cuantifiquen las compensaciones entre principios para alcanzar la opción particular más adecuada. No obstante, hay ocasiones donde la compensación es inaceptable, hay derechos como la dignidad humana que no están sujetos a equilibrio alguno.

1.1.4. Necesidad de especificación y aproximaciones computacionales

La UE, en su Guía para IA confiable, establece diversos requisitos los cuales han de considerarse a la hora de construir una IA confiable (*Figura 1*). Sin embargo, a pesar de que algunos autores muestran aproximaciones metodológicas para entornos de salud (Sáez et al., 2024), existe cierta ambigüedad acerca de qué abarca realmente cada requisito y cómo se puede satisfacer de forma exitosa mediante métodos técnicos.

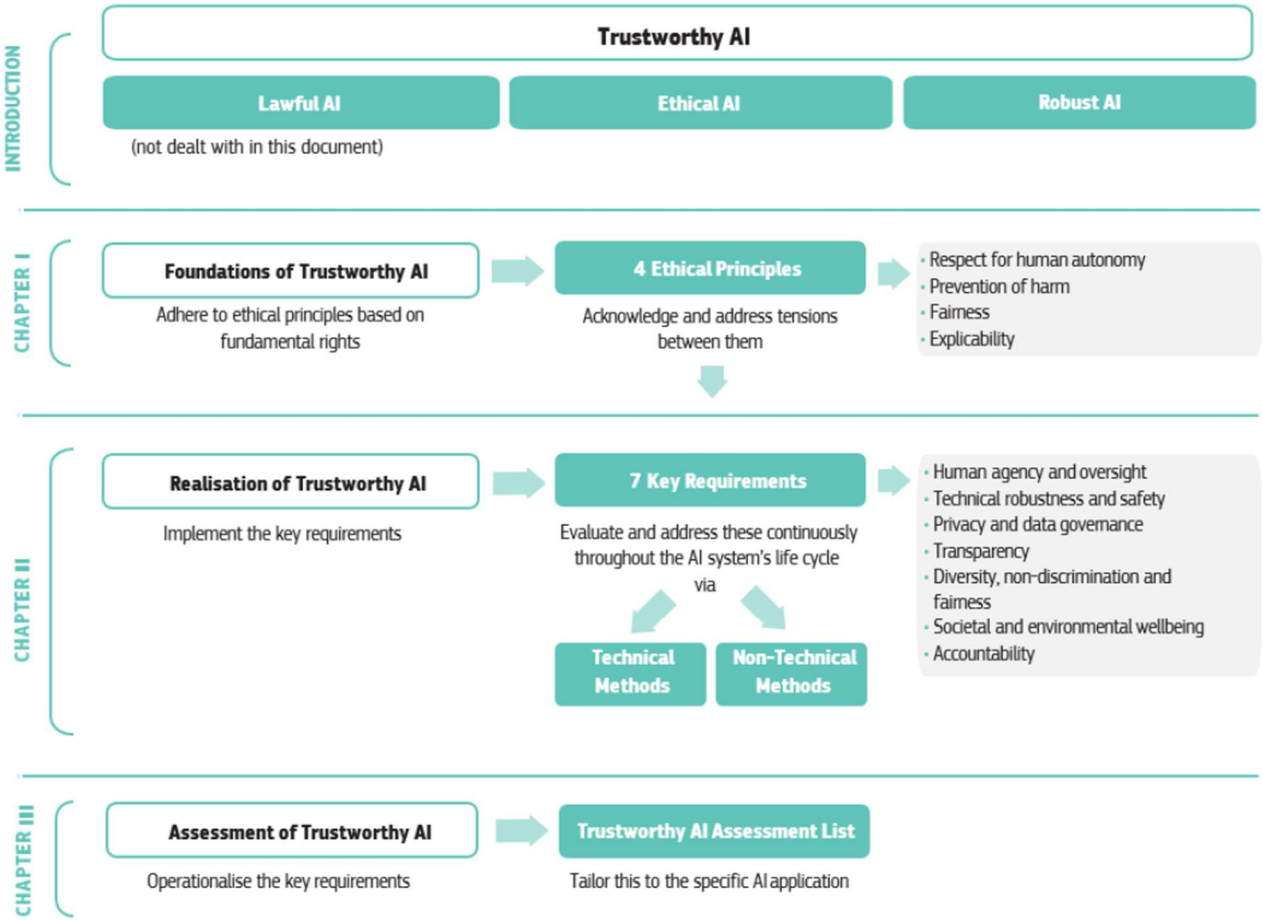

Figura 1. Marco de trabajo para alcanzar una IA confiable (European Comission, 2019).

Dicha ambigüedad, unida a las tensiones entre principios y a la falta de aproximaciones metodológicas técnicas para satisfacer la IA confiable, generan un contexto de falta de determinación donde cobra sentido la investigación y se plantean los objetivos a realizar.

1.2. Objetivos

1.2.1. Objetivo principal

Este proyecto tiene como finalidad **establecer y determinar de forma ejemplificada una guía o *pipeline* (del inglés) para el desarrollo, implementación y gestión tecnológica de una IA confiable en salud**, de modo que los desarrolladores e investigadores de IA en salud puedan contar con una metodología y herramientas para abordar la fiabilidad de un sistema de IA en salud.

1.2.2. Objetivos específicos

El objetivo principal se sustenta en los siguientes objetivos específicos a lo largo del desarrollo del proyecto:

O1. Especificar técnicamente los principios y requisitos de una IA confiable según la Unión Europea relacionados con el ciclo de vida de la IA.

O1.1. Analizar las componentes y condiciones de una IA confiable.

O1.2. Determinar los métodos técnicos necesarios para satisfacer los requisitos especificados enmarcados en el problema de la IA de clasificación.

O1.3. Sintetizar las etapas del ciclo de vida de una IA.

O1.4. Elaborar una matriz que encuadre los métodos técnicos según requisito y etapa en el ciclo de vida.

O2. Generar un pipeline mediante una libreta dinámica que ejemplifique la construcción de una IA confiable.

O2.1. Desarrollar una libreta dinámica de código de programación individual para cada principio.

O2.2. Unificar las libretas con la matriz.

O2.3. Establecer una *checklist* asociada a las etapas de cada libreta que guie la confiabilidad de la IA.

O3. Demostrar la generalización y aplicación de la guía mediante su evaluación en diferentes conjuntos de datos en salud.

O3.1. Buscar conjuntos de datos que puedan ser relevantes para el estudio por la tipología de datos o la sensibilidad de sus variables.

O3.2. Evaluar la efectividad y utilidad de la guía en los conjuntos de datos seleccionados.

Además, atendiendo a los Objetivos de Desarrollo Sostenible (ODS) pertenecientes a la Agenda 2030 establecida por los estados miembros de las Naciones Unidas (United Nations, 2015), la investigación a desarrollar en IA confiable por su directa relación con los principios éticos anteriormente descritos muestra un alto grado de relación con los siguientes ODS:

- **ODS 1.** Fin de la pobreza.
- **ODS 3.** Salud y bienestar.
- **ODS 5.** Igualdad de género.
- **ODS 9.** Industria, innovación e infraestructura.
- **ODS 10.** Reducción de las desigualdades.
- **ODS 16.** Paz, justicia e instituciones sólidas.

CAPÍTULO 2. Antecedentes

Este capítulo tiene como objetivo hacer una revisión del contexto regulatorio, técnico y del estado del arte de modo que, con anterioridad al comienzo de la investigación, se pueda observar cuáles son las tendencias de vanguardia empleadas alrededor del uso y la construcción de una IA confiable, buscando por tanto aportar valor a las propuestas ya existentes.

2.1. Marco europeo de la IA confiable

Comenzamos explorando el marco europeo, ya que es el contexto geográfico al que estamos sujetos y del cual vamos a extraer la información para especificar la metodología. Los principios mencionados en el Capítulo 1 deben traducirse en requisitos concretos que puedan ser aplicados al construir una IA confiable. La siguiente lista proporcionada por la UE en su guía (European Comission, 2019), muestra los requerimientos que han de considerarse, estableciendo igualdad de relevancia entre ellos:

- Privacidad y gobernanza de datos
- Diversidad, no discriminación y justicia
- Transparencia
- Robustez técnica y seguridad
- Agencia humana y supervisión
- Bienestar social y medioambiental
- Responsabilidad

Cada requisito puede estar relacionado con uno o más de los principios éticos, de modo que la totalidad de ellos cubrirán los principios propuestos, y, en consecuencia, los derechos fundamentales propuestos en la Carta de Derechos Fundamentales de la Unión Europea. En la *Figura 2* podemos observar dichas relaciones:

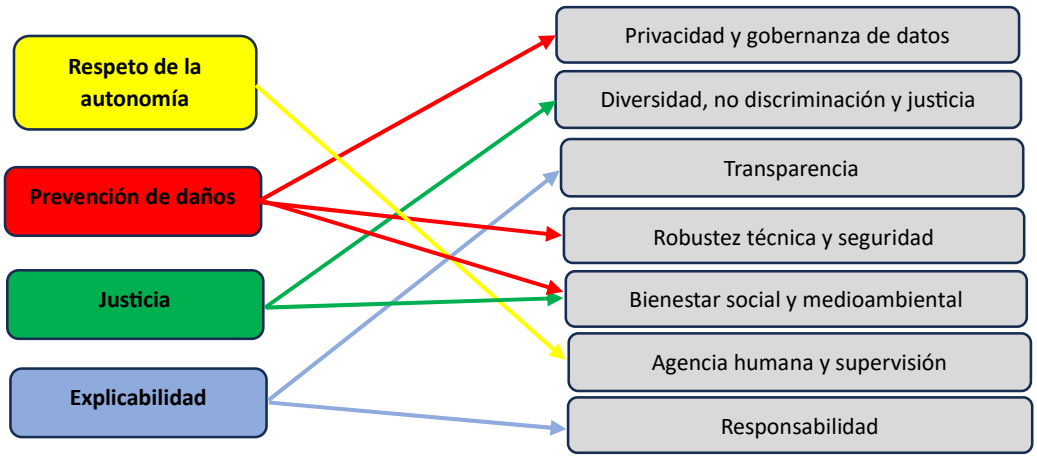

Figura 2. *Principales relaciones entre principios y requisitos de una IA confiable según la guía de la UE*

Para consultar más acerca de las tensiones véase la siguiente bibliografía (Thiebes et al., 2021).

Lógicamente, las relaciones y tensiones entre principios observadas anteriormente se verán reflejadas también en los requisitos, por lo que deberán considerarse las potenciales tensiones y realizar las compensaciones necesarias para el dominio concreto, ya que para una aplicación concreta sí que podría identificarse una jerarquía de relevancia. Pese a no existir ningún tipo de jerarquía, sí que puede ser interesante establecer cierto orden a la hora de cumplir los requisitos. Observamos pues a continuación que componentes involucra cada requisito.

2.1.1. Privacidad y gobernanza de datos

El requisito de Privacidad y gobernanza de datos está ligado con el **principio de prevención de daños.** Es necesario cubrir la calidad e integridad de los datos con una adecuada gobernanza de datos, contar con políticas y protocolos para procesar los datos de modo que se proteja la privacidad. Este requisito será el primer a cumplir, ya que nos asegurará, antes de trabajar con los datos, que estos son de calidad y su tratamiento a lo largo del ciclo de vida será correcto. Comprende los siguientes aspectos:

- **Privacidad y protección de datos:** proteger, tanto toda la información proporcionada inicialmente por el usuario, como la generada durante la interacción de la IA. Además, los datos recogidos no deben utilizarse ilegal o injustamente para discriminar.
- **Calidad e integridad de los datos:** asegurar que los datos contenidos no contengan sesgos o errores. Crucial para el rendimiento del sistema, ya que podrá modificar el comportamiento de la IA. Para ello, cada proceso debe evaluarse y documentarse.
- **Acceso a datos:** considerar protocolos para moderar el acceso a los datos, indicando quién y bajo qué circunstancias tiene acceso. Únicamente debe tener acceso personal apropiadamente cualificado.

2.1.2. Diversidad, no discriminación y justicia

El siguiente requisito se relaciona con el **principio de justicia**. Se debe permitir la inclusión y diversidad de la IA en la totalidad de su ciclo de vida. Hay que considerar tanto a todos los interesados afectados, como asegurar la equidad en el diseño de procesos y tratamiento. Satisfacer este requerimiento en segundo lugar garantizará que, además de contar con datos protegidos y sin errores, estos estén libres de sesgos. Incluye:

- **Evasión de sesgos injustos:** eludir cualquier tipo de inclinación que incite a prejuicios o discriminación de grupos. El sesgo debe eliminarse de todas y cada una de las fases, desde la recolección de datos hasta el desarrollo del sistema de IA, incluyendo procesos de supervisión y formando equipos con multitud de orígenes, culturas y disciplinas que garanticen una diversidad de opiniones.
- **Accesibilidad y diseño universal:** realizar un diseño centrado en el usuario para que, independientemente de las condiciones de la persona (edad, género o habilidad), esta pueda utilizar los servicios. Se debe prestar atención a condiciones más particulares como la discapacidad. El diseño debe considerar los principios de “Universal Design” (UC Berkeley, s. f.) para ampliar el rango de usuarios al que va dirigido y garantizar un acceso equitativo.

- **Participación de los interesados:** mantener durante todo el ciclo de vida de la IA una comunicación con los interesados afectados indirecta o directamente. Puede ser beneficioso solicitar información regularmente incluso una vez implementada la IA para asegurar que esta opera correctamente o advertir posibles mejoras de cara al usuario final.

2.1.3. Transparencia

El requisito de transparencia está estrechamente relacionado con el **principio de explicabilidad**. Abarca los siguientes conceptos:

- **Trazabilidad:** documentar los procesos de modo que pueda identificarse las razones por las que una salida ha sido errónea y, en consecuencia, prevenir futuros fallos. Facilita la auditabilidad y explicabilidad.
- **Explicabilidad:** habilidad de explicar tanto los procesos técnicos de la IA, que estos puedan ser entendidos por humanos, como las decisiones humanas relacionadas. Hay que considerar la compensación con la precisión, el aumento de una suele ir a coste de la otra. Las explicaciones deben adaptarse al tipo de persona. Destacamos también la importancia de diferencia correctamente explicabilidad e interpretabilidad. La primera hace referencia a la posibilidad de explicar la funcionalidad y operaciones de un sistema de manera no técnica a una persona no experta en la materia. Paralelamente, la interpretabilidad requiere un profundo nivel de detalles y ayuda a comunicarse con expertos, estableciendo que es posible para al menos un observador externo entenderlo y encontrar su significado. (Albahri et al., 2023; European Comission, 2020).
- **Comunicación:** los humanos deben tener consciencia de que interactúa con una IA. Estos deben tener la posibilidad de decidir en su contra para cumplir con los derechos humanos. Es importante también comunicar las capacidades y limitaciones de la IA

2.1.4. Robustez técnica y seguridad

Este requisito está también se vincula con el **principio de prevención de daños.** Se debe diseñar con un enfoque preventivo de modo que cualquier daño involuntario o inesperado sea minimizado. Mantener el rendimiento frente a agentes adversos asegurando en cualquier caso la integridad física y mental de las personas.

- **Resiliencia ante ataques y seguridad:** garantizar una protección frente a vulnerabilidades susceptibles de ser atacadas por agentes adversos en cualquier área de la IA: datos, modelo, estructura, o software y hardware. Es bien sabido que ante un ataque el comportamiento puede ser modificado, ocasionando que la IA ofrezca resultados erróneos e incluso cause daños.
- **Plan de emergencia y protección general:** se deben incluir salvaguardas, probadas proactivamente, para minimizar los errores. Las medidas de este plan dependerán del riesgo y la capacidad del sistema.

- **Precisión:** la habilidad de la IA para hacer juicios correctos. Si la evaluación y el desarrollo se realizan de forma correcta se pueden reducir los riesgos de predicciones imprecisas. Además, en aquellos casos donde la imprecisión sea inevitable, la IA debe indicar la probabilidad de los errores. De especial importancia en las situaciones donde el sistema tiene un impacto en la vida humana, como es nuestro caso de la IA en salud.
- **Fiabilidad y reproducibilidad:** la fiabilidad ayuda a prevenir daños inintencionados ya que nos asegura que la IA actúa de forma correcta dentro de un rango de entradas y situaciones. La reproducibilidad es la capacidad de mantener el comportamiento bajo las mismas condiciones. Esto es importante para describir la actuación de una IA y facilitar la evaluación.

2.1.5. Agencia humana y supervisión

Según el **principio de respeto a la autonomía humana**, los sistemas deben apoyar la autonomía humana y la toma de decisiones. Comprende tres conceptos:

- **Derechos fundamentales:** garantizar el cumplimiento de los derechos fundamentales. Se pueden tanto permitir como obstaculizar, por lo que ser una prioridad en la implementación e incluir una evaluación de si los riesgos se pueden reducir o justificar según sea necesario en la sociedad. Los sistemas que potencialmente pueden infringir los derechos humanos deber recibir comentarios externos.
- **Agencia humana:** capacitar a los usuarios para tomar decisiones autónomas informadas. Deben tener información y herramientas para comprender e interactuar satisfactoriamente. La autonomía debe ser central en el funcionamiento del sistema.
- **Supervisión humana:** asegurar que la IA no termina la autonomía humana o causa efectos indeseados. Existen tres tipos: human-in-the-loop (HITL), si el humano interviene en cada decisión durante el ciclo de la IA; human-on-the-loop (HOTL), si interviene durante el diseño y monitoriza la operación; o human-in-command (HIC), HIC si supervisa la actividad en general y decide cuándo y cómo usar la IA en una situación particular. El grado dependerá de la aplicación particular.

2.1.6. Bienestar social y medioambiental

Relacionado con los **principios de justicia y prevención de daños**. Idealmente, las IAs deben beneficiar al medio ambiente, la amplia sociedad (incluidas las generaciones futuras) y otros seres sintientes. Considerar:

- **IA sostenible y respetuosa con el medio ambiente:** analizar los procesos y examinar el uso de recursos y consumo energético durante el entrenamiento, optando por opciones menos dañinas. Garantizar el respeto al entorno en toda la cadena de suministro.
- **Impacto social:** considerar y monitorizar los efectos de la exposición ubicua respecto a la concepción de agencia social. Las IAs pueden tanto incrementar nuestras habilidades sociales como deteriorarlas, afectando al bienestar físico y mental.

- **Sociedad y democracia:** evaluar, más allá del impacto individual, la repercusión de la IA desde una perspectiva social, su efecto en instituciones, la democracia y la sociedad en general. Es de particular importancia en situaciones relacionadas con procesos democráticos.

2.1.7. Responsabilidad

El requisito de transparencia está estrechamente relacionado con el **principio de explicabilidad**. Abarca los siguientes conceptos:

- **Auditabilidad:** habilitar la evaluación de algoritmos, datos y procesos de diseño. No implica que la información del modelo de negocio o propiedad intelectual estén abiertamente disponible. La evaluación por auditores internos y externos puede contribuir a la confiabilidad.
- **Minimización y notificación de impactos negativos:** permitir la capacidad de reportar acciones o decisiones y responder a las consecuencias de una salida. Identificar, evaluar y documentar los impactos negativos. Realizar evaluaciones proporcionales al riesgo
- **Compensaciones:** valorar las tensiones entre los requisitos implementados. Deben dirigirse de un modo racional y metódico dentro del estado del arte. Identificar situaciones donde las compensaciones no son éticamente aceptables. El que toma las decisiones es responsable de la compensación y debe revisar la salida para asegurar los cambios necesarios para el sistema.
- **Ajustes:** asegurar una rectificación adecuada cuando ocurre un impacto adverso. Para la fiabilidad, es clave conocer la posibilidad del reajuste cuando las cosas van mal.

La existencia de todos estos requisitos ha conllevado la necesidad de definir una regulación capaz de limitar y abordar desde un punto legal y menos laxo la construcción de una IA confiable. Los elementos anteriores componen una simple guía elaborada por la UE. En ningún caso abarca obligaciones legales, ya que el objetivo del documento es abarcar únicamente las componentes ética y robusta de la IA, Los límites legales se establecen a través de una regulación legislativa

2.2. Regulación de la IA

Con el fin de estructurar los requisitos dentro de un marco jurídico, el Parlamento Europeo y el Consejo de la UE alcanzaron un acuerdo político sobre la Ley de IA el pasado abril de 2024 (European Parliament, 2024), situando además al continente como pionero en la regulación para abordar los riesgos asociados a la IA.

La Ley de IA tiene entre sus objetivos establecer limitaciones y controles de riesgos para los requisitos de una IA confiable para los desarrolladores e implementadores de sistema, haciendo además hincapié en los contextos donde el riesgo es considerablemente elevado. En las normas propuestas se prohíben prácticas inaceptables y se establecen requisitos claros para los sistemas de alto riesgo, evaluados mediante una jerarquía de 4 niveles: inaceptable, alto, limitado y mínimo.

En la legislación, se describe mediante una secuencia de pasos la declaración de conformidad de los proveedores de IA de alto riesgo, incluyendo también las medidas de seguimiento y supervisión que deben emplearse tras eldatas despliegue.

A pesar de que en la legislación se hace mención de los requisitos propuestos en la guía europea, tan solo establecen limitaciones y en ningún caso orienta de forma práctica la realización de un sistema. El foco reside en los componentes que, a nivel general, debe satisfacer un sistema de IA de alto riesgo. Además, los límites propuestos en la legislación son en gran medida ambiguos, ya que se hace un amplio uso de adjetivos como "apropiado" donde, sin un contexto de referencia, existe confusión acerca de cuál es el nivel buscado.

2.3. Ciclo de vida de la IA

Los requisitos definidos por la UE para una IA confiable deben satisfacerse a lo largo de todo el ciclo de vida del sistema, para lo cual es necesario saber pues cuales son aquellas fases que componen un ciclo de vida. Actualmente, parece no existir ningún tipo de consenso acerca de las fases que componen el desarrollo de la IA, por lo que el objetivo de esta sección es analizar aquellos estándares más consolidados con el fin de escoger un modelo o conjunto de varios que se ajuste en mayor medida a nuestra aplicación. Así pues, exponemos algunos de ellos de manera objetiva para, en el Capítulo 3, determinar cuál será nuestro ciclo de vida concreto.

2.3.1. CRISP-DM (Cross-Industry Standard Process for Data Mining)

CRISP-DM es el Proceso Estándar Cruzado para la Minería de Datos. Proporciona un marco estructurado para proyectos vinculados con la ciencia de datos, siendo a su vez el ciclo de vida más común para los proyectos de esta índole (J. S. Saltz, 2021; *SPSS Modeler Subscription*, 2021). En este estándar se definen 6 fases principales (*Véase Figura 3*):

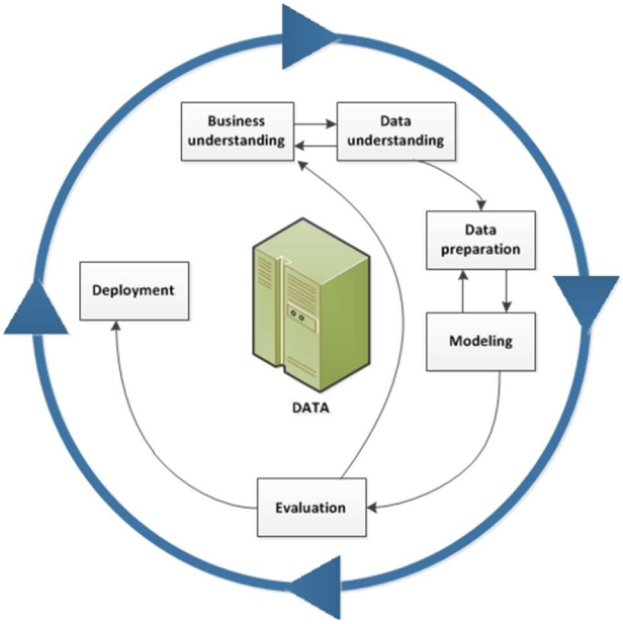

Figura 3. *Ciclo de vida de minería de datos según CRISP-DM (*SPSS Modeler Subscription*, 2021).*

- **Comprensión del negocio:** determinar, desde un punto de vista empresarial, los objetivos de la minería de datos, así como evaluar la situación y los recursos del proyecto.

- **Comprensión de los datos:** identificar y analizar los datos que servirán como soporte para los objetivos del proyecto. En esta fase se recopilan los datos iniciales al tiempo que se realiza un control de su calidad y un análisis exploratorio de ellos.
- **Preparación de datos:** limpiar los datos (lidiar con datos perdidos, inconsistencias, etcétera) y seleccionar las características relevantes.
- **Modelado:** construir y evaluar los modelos. En esta fase se forman los diseños de prueba y se estima su rendimiento.
- **Evaluación:** evaluar qué modelo satisface en mayor medida los objetivos empresariales. Se revisa el proyecto para identificar áreas de mejor y se decide si el modelo está listo para implementarse o si, por el contrario, se han de realizar más iteraciones para mejorar el proyecto.

Cabe destacar que algunos autores (García-Gómez et al., 2019) han propuesto adaptaciones de CRISP-DM para el ámbito de IA en salud. Además existen otros estudios acerca del ciclo de vida de una IA (Ng et al., 2022) en salud aunque, hasta nuestro conocimiento, sin una consolidación comparable a las metodologías expuestas.

2.3.2 GenAI (Generative AI) Life Cycle

Este ciclo de vida está confeccionado para definir los pasos de construcción de aplicaciones basadas en IA generativa, como pueden ser *chatbots* o asistentes virtuales (J. Saltz, 2024). Este tipo de sistemas se caracterizar por ser capaces de crear contenido como texto o imágenes. Tal y como podemos apreciar en la *Figura 4* ciclo de vida típico consta de las siguientes fases:

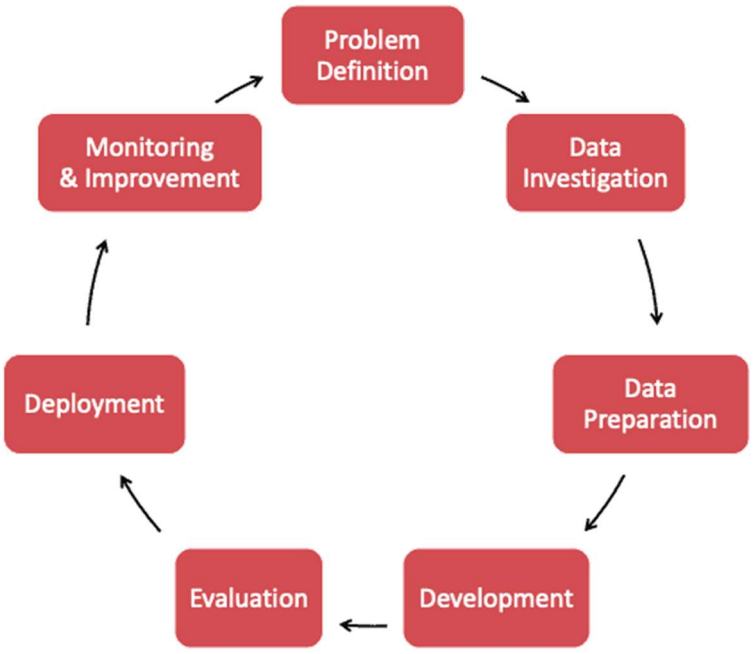

Figura 4. *Ciclo de vida de una IA generativa (GenAI)(J. Saltz, 2024).*

- **Definición del problema:** establecer los objetivos del proyecto desde un punto de vista comercial, identificando el desafío que debe abordar la GenAI.
- **Investigación de datos:** situar el foco en evaluar el contexto de los datos: disponibilidad, relevancia, calidad, etcétera.
- **Preparación de datos:** realizar una limpieza del conjunto de datos para adecuarlos al modelo a emplear.
- **Desarrollo:** construir la aplicación GenAI empleando modelos que procesen texto de forma similar al lenguaje humano, Large Language Model (LLM), y complementando con Generación Aumentada de Recuperación (RAG), una modalidad que permite a los sistemas acceder y emplear información externa a la hora de generar texto.
- **Evaluación:** realizar pruebas para garantizar el rendimiento y confiabilidad de la aplicación. La evaluación se efectúa acorde con unos criterios vinculados con los estándares y las necesidades comerciales particulares.
- **Despliegue:** implementar la aplicación en el contexto particular. Se configura la infraestructura a emplear, así como el establecimiento de un protocolo de seguimiento para monitorizar la evolución del modelo de acuerdo con su rendimiento y la información de los usuarios finales.

2.3.3. AWS Well-Architected Framework

La siguiente arquitectura corresponde a un marco de trabajo cíclico e iterativo donde se definen las instrucciones y mejores prácticas para desarrollar un proyecto de aprendizaje automático en AWS (Amazon Web Services) (*Well-Architected machine learning lifecycle - Machine Learning Lens*, 2023). AWS es una plataforma en la nube que ofrece diferentes servicios de bases de datos, análisis o IA. Es ampliamente utilizado por desarrolladores y organizaciones por si fiabilidad y accesibilidad. En la *Figura 4* referenciada se determinan las siguientes fases del ciclo de vida de una IA:

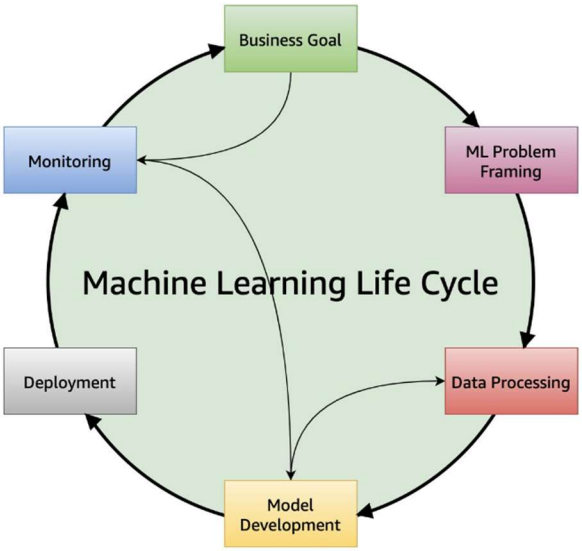

Figura 5. *Ciclo de vida de un proyecto de machine learning en Amazon Web Services (AWS)(*Well-Architected machine learning lifecycle - Machine Learning Lens*, 2023).*

- **Identificación de los objetivos empresariales:** definir el problema empresarial y el valor obtenido tras la resolución de la problemática.
- **Formulación del problema de ML:** traducir el problema empresarial en problema de aprendizaje automático, determinando lo que se busca predecir y cómo se optimizará el rendimiento del sistema.
- **Procesamiento de datos:** abarca tanto la recopilación y preparación de los datos, como procesos de selección de características, de tal modo que se confeccionen los datos para ser directamente utilizados por el modelo.
- **Desarrollo del modelo:** construir, entrenar y evaluar el modelo. Se integra dentro de esta fase la creación de un pipeline que automatice los procesos a realizar.
- **Implementación:** desplegar en el contexto específico, tras el entrenamiento y la evaluación, para que realice las predicciones necesarias.
- **Monitorización:** realizar un seguimiento del comportamiento del modelo, de modo que mantenga un nivel de rendimiento y se pueda hacer frente con los potenciales problemas existentes.

2.3.4. MLOps

El esquema que se presenta a continuación se basa en un enfoque cíclico e iterativo, estableciendo las directrices y mejores prácticas necesarias para llevar a cabo un proyecto de aprendizaje automático conforme a los principios de MLOps (Machine Learning Operations). MLOps es un conjunto de prácticas que combina Aprendizaje Máquina (ML, del inglés Machine Learning), DevOps y Data Engineering para automatizar y mejorar la gestión del ciclo de vida del aprendizaje automático (Neupane, 2023; *Understanding MLops Lifecycle*, 2024). Es ampliamente adoptado por desarrolladores y organizaciones por su capacidad de aumentar la eficiencia y la fiabilidad de los proyectos de ML. En la *Figura 5* observamos como el ciclo de vida de una IA en el contexto de MLOps se compone de tres fases principales:

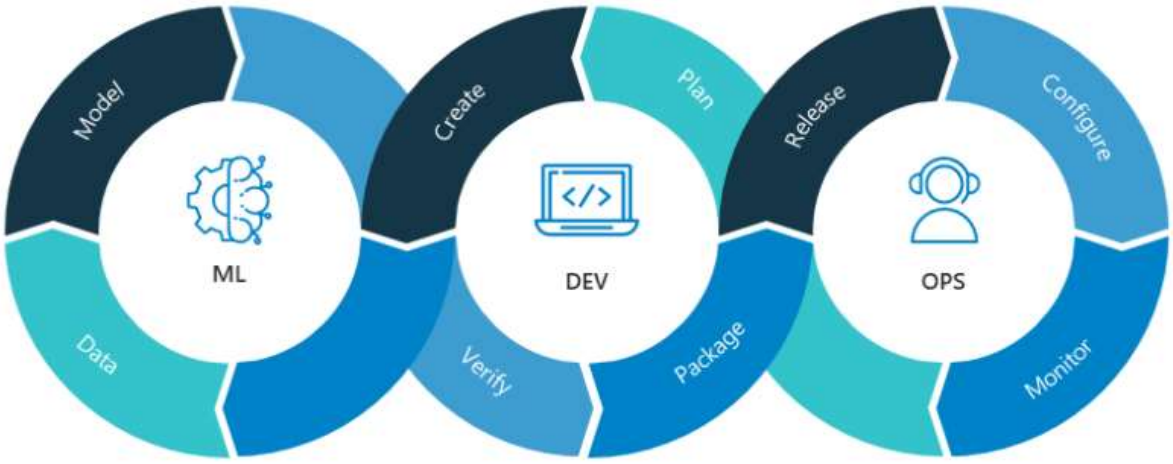

Figura 6. Ciclo de vida de un sistema de aprendizaje máquina según MLOps (Neupane, 2023).

- **Aprendizaje Máquina (ML):** incluye tanto recopilar datos relevantes de diversas fuentes como limpiar los datos para adaptarlos al análisis. En esta etapa también se desarrollan y entrenan los modelos de ML.
- **Despliegue (Dev):** integrar el modelo con las aplicaciones existentes y configurarlo con la infraestructura necesario, todo ello tras el entrenamiento y la evaluación.
- **Operaciones (Ops):** supervisar continuamente el rendimiento del modelo para determinar cualquier variación y realizar los ajustes necesarios para que se mantenga el rendimiento. En esta etapa también se gestiona la infraestructura subyacente al modelo, las versiones de los datos y los modelos, y se realizan procesos para automatizar al máximo el pipeline.

2.4. Aproximaciones relacionadas

Previo a la construcción de la metodología o pipeline investigamos acerca de la existencia de algún tipo de aproximación que incorpore métodos técnicos para lidiar con la elaboración de un sistema de IA confiable.

A nivel general, tal y como hemos explicado anteriormente, encontramos la guía propuesta por la UE (European Comission, 2019), donde la información más útil desde el punto de vista práctico de un desarrollador es la extraíble de la 'Trustworthy AI Assesment List' ofrecida. En este extenso documento se plantean diferentes preguntas que, como profesional en la construcción de un sistema, deberías considerar. Estos enunciados responden a los componentes particulares existentes en cada requisito, formulando preguntas de respuesta abierta que difícilmente pueden evaluarse de forma objetiva: "¿Consideras desarrollar la IA con el mínimo uso de datos potencialmente sensibles?", "¿Consideras vías para medir si tu sistema realiza una cantidad inaceptable de predicciones imprecisas?". Respecto a preguntas de esta índole, es necesario determinar conceptos como datos sensibles, qué es una predicción imprecisa, o qué cantidad de predicciones sería un umbral inaceptable. Así pues, idealmente se deberían de ofrecer métodos técnicos y cuantitativos que permitan evaluar y determinar de forma cuantitativa el cumplimiento de los requisitos.

La consultora Deloitte propone su 'Trustworthy AI Framework' basado en el ciclo de vida MLOps (Deloitte, 2022). En la misma sección web donde ofrece el marco de trabajo, también opina acerca la IA confiable, destacando su importancia y practicidad en el mundo actual. Visualizando su portal web oficial, en él se ofrece una matriz circular con los requisitos y los componentes individuales dentro de ellos, realizando una aproximación similar a la de la UE, ya que nuevamente se prescinde de la utilización de métodos técnicos.

Del mismo modo, la empresa IBM también muestra interés en la IA confiable y ofrece en su página web (IBM, 2021) un apartado dedicado exclusivamente a esta. En esta sección, a pesar de no ofrecer una solución holística para la construcción de una IA confiable, se hace una recopilación de estudios vinculados a cuestiones relacionadas con la confiabilidad de un sistema, como es el caso de la explicabilidad o la cuantificación de la incertidumbre.

Observando propuestas más firmes y desarrolladas, encontramos el artículo "*A responsible AI framework: pipeline contextualisation*" (Vyhmeister et al., 2023). En esta publicación se consigue ir más allá y, a través de un conjunto de preguntas a modo de *checklist*, se realiza un flujograma completo. El esquema cuenta con múltiples ramificaciones que definen las características de tu sistema en base, principalmente, al riesgo que compone tu sistema de IA. La metodología es mucho más completa, aunque en ningún caso se estudian o proponen métodos computacionales a partir de los cuales construir tu modelo para posteriormente evaluarlo con su 'pipeline'.

Podemos observar la existencia de estudios sistemáticos, incluso centrados en salud, y situándose por tanto muy cerca de nuestro ámbito de aplicación. Así pues, el artículo "*A systematic review of trustworthy and explainable artificial intelligence in healthcare: Assessment of quality, bias risk, and data fusion*" (Albahri et al., 2023) proporciona un marco interesante para investigar acerca de los sistemas de apoyo a la decisión en salud. Sin embargo, su alcance corresponde a la determinación de métodos actuales útiles para el desarrollo de IAs confiables y explicables, por lo que en ningún caso ofrece algún tipo de metodología completa a partir de la cual construir tu sistema.

CAPÍTULO 3. Materiales y métodos

Durante este capítulo se desglosará la metodología empleada a modo de mapa genérico que sirva de guía de la investigación. Además, se presentarán todos aquellos materiales y métodos empleados para la realización del estudio y se analizarán los conjuntos de datos empleados, así como las herramientas que se han aplicado para manejarlos.

3.1. Metodología

La metodología presentada a continuación es fruto de un proceso de diseño orientado a abordar la investigación y desarrollo de nuestra solución para el desarrollo de una IA confiable. El flujo de trabajo se ha estructurado con el fin de proporcionar un marco claro y sistemático de forma que podamos garantizar calidad y rigurosidad a la hora de abordar la investigación.

Dentro de lo que es la IA y en concreto el ML, encontramos una rama predictiva compuesta por problemas de clasificación, con salidas categóricas o discretas (Ej. reingreso o no reingreso), y problemas de regresión con salidas numéricas o continuas (Ej. número de meses de supervivencia de un paciente). Si bien la mayoría de métodos técnicos son extrapolables a múltiples enfoques de ML, en el presente estudio nos focalizaremos en problemas de clasificación, dejando la regresión para trabajo futuro.

La *Figura 7* presenta mediante un diagrama de flujo la metodología empleada en este trabajo mediante la secuencia lógica de pasos que guiarán el desarrollo de la investigación.

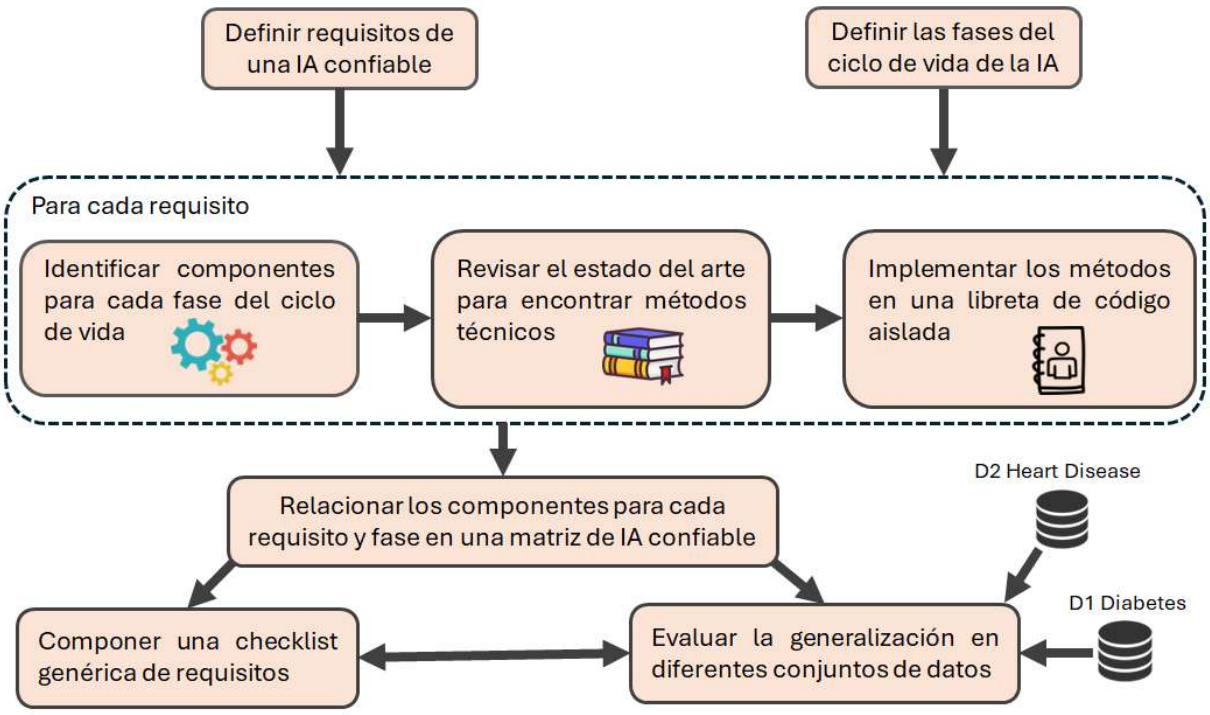

Figura 7. *Flujograma de la metodología de trabajo.*

3.2. Herramientas

3.2.1. Entorno de trabajo

El estudio se ha llevado a cabo combinando sinérgicamente libretas del lenguaje de programación Python en Jupyter Notebook (en sus versiones 3.12.1 y 7.0.7 respectivamente), un entorno que permite combinar código con texto de un modo eficiente. Las libretas se han integrado en el entorno de desarrollo Visual Studio Code (VS Code), en su versión 1.84. Python, como lenguaje de programación base en el proyecto, destaca por su versatilidad y robustez en la vanguardia de la investigación. Cuenta con una sintaxis clara y concisa que, sumada a la amplia variedad de bibliotecas especializadas, lo convierten en una herramienta extremadamente atractiva para el análisis de datos y el desarrollo de modelos de aprendizaje automático.

El auge continuo de Python durante los últimos años ha hecho de este un lenguaje central en cualquier estudio de IA, permitiendo a su vez que el acceso a métodos y algoritmos se encuentre completamente actualizado, representando el estado del arte en múltiples áreas científicas.

Particularmente, el uso de libretas o 'notebooks' ha sido la opción adoptada como método preferido en contraposición a archivos compuestos exclusivamente por código. La elección se apoya en la necesidad de proporcionar de manera extensa y clara una ejemplificación a modo de tutorial con capacidad ejecutable del proceso de desarrollo de la IA confiable. Es necesario pues realizar una estructura jerarquizada por secciones e incluir texto explicativo complementario de modo que se facilite la comprensión y se guie al usuario durante el flujo de trabajo. En este sentido, prescindir de las libretas podría ser origen de confusión y desorientación entre los posibles desarrolladores finales a causa de la falta de claridad y contexto que puede existir en archivos de código puro.

Así pues, el entorno combinado de Jupyter Notebook y VS Code supone una base plataforma flexible y extremadamente potente para maximizar la eficiencia y organización en el desarrollo del modelo sin prescindir de la estructura y claridad necesarias para su comprensión

Además, cabe mencionar que para realizar un control de versiones se han empleado Git y GitHub. Git es un sistema de control de versiones que utiliza la plataforma de desarrollo colaborativo GitHub. Este servicio es muy interesante, ya que permite colaborar en proyectos de software desde diferentes ubicaciones, controlar el acceso de estos proyectos y documentarlos de forma adecuada.

3.2.2. Modelos base

En cuanto a los algoritmos utilizados, destacamos principalmente dos: Random Forest y Naive Bayes. Random Forest corresponde a aproximación más compleja y completa para abordar problemas de aprendizaje automático, mientras que Naive Bayes supone una ejemplificación de uso de modelos más sencillos.

- **Random Forest:** es una técnica de aprendizaje automático basada en la construcción de múltiples árboles de decisión. El modelo combina la predicción de todos ellos para obtener un resultado más robusto y preciso. Cada árbol se construye utilizando una muestra aleatoria de entrenamiento y seleccionando, también aleatoriamente, características en cada división del árbol. La predicción se hará en base a la clase más común al final de la rama y mediante el

consenso de todos los árboles la predicción final. Random Forest emplea múltiples clasificadores que individualmente son débiles pero que cuando se emplean en conjunto hacen de la técnica un método altamente resistente al sobreajuste y muy adecuado para datos con gran cantidad y tipos de características y clases.

- **Naive Bayes:** es un algoritmo de clasificación basado en el teorema de Bayes. Este modelo asume que las características son independientes entre sí, lo que simplifica el cálculo condicionado de probabilidad. En nuestro caso, para generalizar su uso, combinamos los modelos 'GaussianNB' y 'CategoricalNB' de Python. GaussianNB se emplea para modelar las características numéricas, asumiendo que estas siguen una distribución gaussiana. Por otra parte, CategoricalNB se empleará para el manejo de las categóricas. Tal y como se entiende la definición del modelo Naive Bayes, la probabilidad final se calculará mediante el productorio de todas las probabilidades condicionadas, por lo que el resultado de la clasificación se obtiene a partir del productorio de ambos modelos por separado. Gracias a su sencillez, la velocidad de entrenamiento y predicción es generalmente rápida, por lo que es muy útil para grandes volúmenes de datos. Sin embargo, su rendimiento puede verse afectado si no se cumplen las suposiciones de independencia.

3.3. Conjuntos de datos

3.3.1. Conjunto de datos "Diabetes 130-US Hospitals for Years 1999-2008"

El conjunto de datos "Diabetes 130-US Hospitals for Years 1999-2008", al cual nos referiremos a partir de este punto mediante la simplificación "Conjunto Diabetes", conforma el grupo principal de datos en el que se evaluará la propuesta tecnológica desarrollada. La información de este conjunto de datos es el producto resultante de una recopilación a lo largo de casi 10 años de 101768 instancias de pacientes diabéticos provenientes de 130 hospitales estadounidenses (John Clore, 2014; Strack et al., 2014). El volumen de datos y el origen de ellos hace de éste un conjunto de datos especialmente interesante, ya que ayuda refleja un contexto realista.

Este conjunto de datos está formado por 52 columnas, incluyendo categóricas nominales y ordinales, numéricas y binarias, combinando a su vez variables de tipo texto con números y categorías codificadas. Entre las características encontramos todo tipo de variables médicas como el tipo de admisión ('admission_type_id') o el diagnóstico primario ('diag_1'), codificado según la Clasificación Internacional de Enfermedades en su versión 9 (CIE-9) (World Health Organization, 2024). También aparecen registradas variables demográficas como la raza o el sexo. El conjunto de datos comprende un problema de clasificación, donde la última columna 'readmitted' corresponde a la etiqueta a predecir, donde se busca predecir aquellos pacientes que no readmiten (NO), los que lo hacen en un periodo inferior a 30 días (<30) o en uno superior a 30 días (>30). La relevancia de estos datos reside en el hecho de que, aquellos pacientes que reingresan en <30 días, muestran una morbimortalidad asociada a la enfermedad superior (Ostling et al., 2017). Así pues, en este trabajo se buscaría conocer aquellos casos donde se prevé un reingreso hospitalario en <30 días puesto que en ese caso no debería de formalizarse el alta.

Cabe mencionar, que, según apuntan algunos estudios realizados con el presente conjunto de datos (Shang et al., 2021), el rendimiento de los modelos predictivos para clasificar el tipo de reingreso no parece ser elevado. Particularmente, para un clasificador Random Forest, el anterior estudio ofrece unos resultados de área bajo la curva ROC de 0.66.

3.3.2. Conjunto de datos "Heart Disease"

El conjunto de datos "Heart Disease", al cual nos referiremos a partir de este punto mediante la simplificación "Conjunto Heart Disease", constituye una muestra más simplificada que el "Conjunto Diabetes" para comprobar la generalización de la propuesta desarrollada para una IA confiable. El conjunto de datos es ampliamente utilizado como *benchmark* en la IA en salud. Tiene su fecha de origen en 1988 y se forma a través de bases de datos de Cleveland y Hungría (Alizadehsani et al., 2019). En este caso, el conjunto de datos cuenta con 11 características y 1190 encuentros.

El conjunto tiene como etiqueta a predecir la presencia de enfermedad cardíaca (1 si existe, 0 si es un paciente normal). Para ello, contamos con 5 variables numéricas y 6 categóricas. De estas últimas cabe decir que tienen un volumen de categorías muy inferior al conjunto Diabetes.

Del mismo modo, analizando el rendimiento de estudios predictivos, podemos esperar unas métricas superiores al "Conjunto Diabetes" tras la evaluación de metodología (Ali et al., 2021). En concreto, en el estudio anterior se emplean métodos de aprendizaje supervisado que permiten alcanzar métricas de área bajo la curva ROC superiores a 0.9.

CAPÍTULO 4. Resultados

El Capítulo 4 tiene como objetivo mostrar los resultados de la metodología a nivel general, explicando los conceptos teóricos y el flujo de trabajo ideal, así como plasmar la evaluación de la metodología en los dos conjuntos de datos explicados anteriormente: “Conjunto Diabetes" y “Conjunto Heart Disease”.

En favor de la ciencia abierta, todos los resultados siguientes así como el código fuente se encuentran accesibles públicamente en el repositorio GitHub: https://github.com/bdslab-upv/trustworthy-ai.

4.1. Matriz de IA confiable

El primer resultado aportado es el desarrollo de una matriz de IA confiable que relaciona los requisitos a lo largo del ciclo de vida con el objetivo guiar a los desarrolladores de IA en la construcción de modelos de IA confiable. Para abordar el desarrollo de un sistema de este tipo, debemos tener en cuenta todos los requisitos que esta ha de cumplir, así como todas las etapas que componen su ciclo de vida.

En primer lugar y tras revisar las metodologías presentadas en el Capítulo 2, proponemos la siguiente estructura para las fases del ciclo de vida de una IA confiable:

- **Preparación de datos:** esta etapa implica la preparación de los datos necesario para realizar el entrenamiento y las pruebas de evaluación del modelo de IA. A su vez, se subdivide en una fase de preprocesado de datos, donde se pretende realizar una limpieza, transformación, y selección de características, así como otra fase de control de calidad de los datos, donde se evaluará la confianza y robustez técnica de estos para la aplicación en cuestión
- **Desarrollo del Modelo:** durante esta etapa se realizará el diseño, entrenamiento y validación de los modelos. Así pues, se comenzará con una fase de diseño del modelo, donde se deberá de escoger a rasgos generales el tipo de modelo predictivo según el tipo de datos (Ej. Regresores, basados en árboles, etcétera) así como los mejores parámetros para este teniendo en cuenta los principios éticos y de robustez técnica y. Además se realizará un entrenamiento para posteriormente evaluar el rendimiento de forma que las métricas obtenidas sean lo más representativas y precisas posibles.
- **Despliegue y uso:** una vez desarrollado y evaluado el modelo, debemos estudiar su implementación en un contexto real. Para ello, en una primera fase debemos evaluar las predicciones del modelo en situaciones del mundo real, y en una segunda, observar cómo se comporta frente a situaciones inesperadas o nuevos datos. En esta etapa también es crucial monitorizar continuamente el rendimiento del modelo y ajustarlo según sea necesario para que continue satisfaciendo los principios éticos.
- **Gestión:** esta última etapa tiene como fin documentar, auditar y estudiar los resultados del modelo. La gestión implica una fase realizar registros detallados del modelo, explicando cómo se ha creado y entrenado, así como incorporando cualquier tipo de cambio o ajuste realizado a lo largo del tiempo. En esta etapa también se deben incluir auditorías periódicas para asegurar el buen funcionamiento e intercambios y colaboración con las partes interesadas y los usuarios finales.

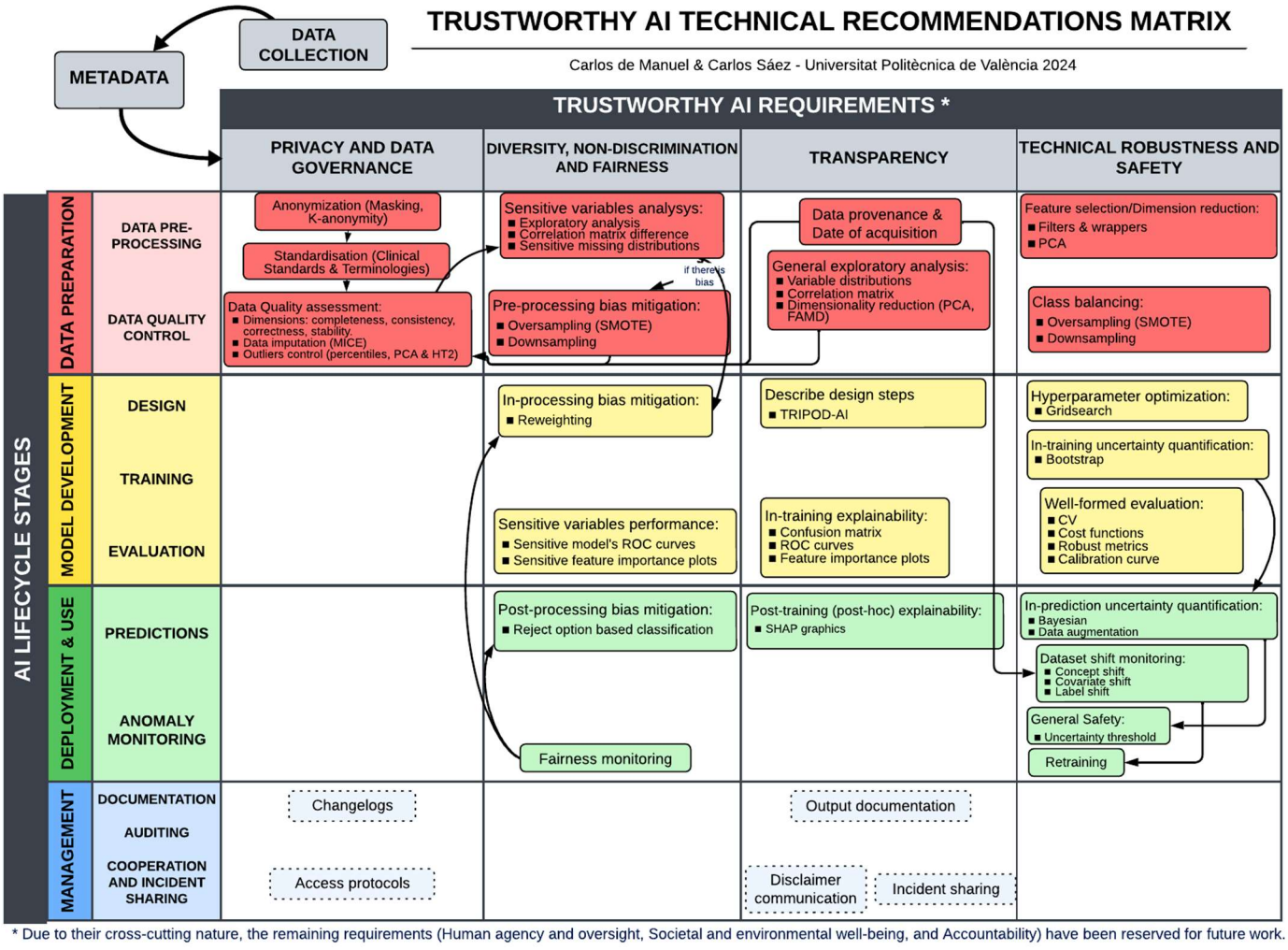

Figura 8. Matriz de requisitos para una IA confiable según la etapa del ciclo de vida.

Seguidamente, si tomamos el marco de referencia de la UE, los requisitos están establecidos por la propia Comisión Europea (European Comission, 2019). Por lo que, una vez definidas las etapas del ciclo de vida de un sistema, podemos estructurar una matriz que contenga recomendaciones técnicas a realizar para satisfacer los requisitos a lo largo de la construcción del sistema. En la *Tabla 1* apreciamos el resultado final de la matriz.

Tal y como se puede observar, la matriz tiene únicamente cumplimentadas 4 columnas, ya que, a diferencia del resto, las tres restantes (Agencia humana y supervisión, Bienestar social y medioambiental, y Responsabilidad) no tienen una resolución técnica directa e independiente. Aunque consideramos como trabajo futuro el tratamiento de estos tres requisitos, es cierto que podrían tratarse dentro de otros, como la Agencia humana y supervisión dentro de Diversidad, no discriminación y justicia debido a su implicación con los derechos fundamentales, o la Responsabilidad que se relaciona con la explicabilidad de Transparencia y la auditabilidad realizada en Robustez y seguridad, donde el sistema realizaría una advertencia que motive al reentrenamiento en caso de observarse un descenso significativo del rendimiento.

Observamos que previo a la entrada de la matriz tenemos una etapa previa denominada 'Metadatos' dirigida al acondicionamiento del conjunto de datos y al aporte de información complementaria útil para el estudio.

Además, pese a que la UE no establece ningún tipo de jerarquía, nosotros proponemos el orden de resolución según la ordenación de las columnas. Es decir, comenzar con **Privacidad y gobernanza de datos**, de modo que nos aseguremos que el material con el que vamos a trabajar es el adecuado; **Diversidad, no discriminación y justicia**, para evitar la realización de sesgos durante el entrenamiento o, en caso de haberlos, detectarlos y cuantificarlos; **Transparencia**, para poder entender de manera completa tanto el tipo de datos que tenemos como su distribución y posibles relaciones, así como también desarrollar herramientas complementarias para comprender las salidas; y por último **Robustez y seguridad**, para asegurar que el rendimiento de nuestro modelo es máximo y está capacitado para lidiar con incertidumbres.

4.2. Pipelines software para IA confiable

A continuación, con el fin de analizar en mayor profundidad la matriz, desglosamos de forma particular cada requisito. Cabe indicar que en cada sección se tendrá acceso a un anexo externo, donde a través de un hipervínculo se podrá observar una libreta demostrativa compuesta por el código y los resultados de la ejemplificación de su uso para los respectivos conjuntos de datos.

A modo de consideraciones previas, exponemos la importancia de usar funciones externas o 'pipelines' para simplificar el código. En nuestro caso, se ha elaborado de forma aislada el modelo mixto de *Naive Bayes*, así como una clase 'handleData' cuyo fin es proporcionar métodos para codificar y decodificar datos, agrupar características, perturbar muestras u obtener elementos aleatorios del conjunto.

4.2.1. Metadatos

Previo a la entrada de la matriz tenemos la etapa de recolección de datos, la cual abarca todos los aspectos relacionados con la obtención de ellos. El paso de esta etapa al comienzo de la guía se hace por medio de una fase intermedia donde se preparan los metadatos. Esta fase tiene como objetivo preparar la información para el 'pipeline' desarrollado. Así pues, en ella deberemos ajustar el formato de los propios datos e incluir información adicional y complementaria que, posteriormente, será utilizada para satisfacer los diferentes requisitos. Los metadatos que deben incluirse son los siguientes, donde señalamos entre paréntesis el nombre empleado en el código:

- **Datos (dataset):** conjunto de datos con todas las características y etiqueta o clase a predecir.
- **Salida (output):** nombre de la columna de la clase a predecir.
- **Clase positiva (positive_class):** en caso de haber más de una clase en la variable a predecir, cuál sería la clase positiva o de mayor peso. Empleada para establecer una jerarquía en la clasificación o binarizar el problema en caso de ser necesario o desearlo.
- **Características identificativas (feat_id):** nombre de la variable o variables empleadas para identificar y diferenciar de forma única cada instancia, ya sea de forma codificada o mediante datos demográficos (Ej. nombre, edad…).

- **Características sensibles (feat_sensitive):** nombre de las variables caracterizadas como sensibles, susceptibles de generar sesgos en los derechos fundamentales de las personas a las cuales la IA aplica. De forma general, la UE caracteriza como sensibles los datos de: origen racial o étnico, opiniones políticas y creencias religiosas o filosóficas, genéticos, relacionados con la salud, y respectivos a la vida u orientación sexual (Parlamento Europeo, 2016).
- **Tipo de variables (feat_types):** diccionario con el nombre y tipo de variable. Escoger entre los 2 grupos generales 'numérica' o 'categórica'.
- **Característica para balancear (feat2balance**): nombre de la variable sensible de la cual se intentará eliminar el potencial sesgo o discriminación por desbalanceo de instancias.
- **Procedencia de los datos (data_provenance):** lista con dos variables de texto, la primera entrada contiene información general acerca del linaje y procedencia de los datos; en la segunda entrada se encuentra el nombre de la variable que da información del linaje para cada instancia.
- **Fecha de adquisición (acquisition_date):** lista con dos variables de texto, la primera entrada con información general del conjunto de datos y la segunda con el nombre variable que define la fecha en la que se adquirió cada instancia. Esta información puede ser usada con propósitos de evaluación temporal del modelo IA o aprendizaje continuo.

Una vez recopilados los metadatos y adaptado al conjunto de datos, se almacenarán exportándose desde Python a un archivo JSON de modo que pueda accederse desde cada requisito cómodamente. Seguidamente, puede comenzarse a abordar la construcción de la IA confiable mediante la metodología establecida que, tal y como proponemos, comienza en la Privacidad y gobernanza de datos.

4.2.2. Privacidad y gobernanza de datos

La sección de Privacidad y gobernanza de datos va encaminada a cubrir principalmente la calidad e integridad de los datos. Además, también abarca temas relacionados con la privacidad, como la anonimización para cumplir con la normativa de protección de datos (Parlamento Europeo, 2016), y metodologías de gobernanza y estandarización . Así pues, las acciones a realizar estarán concentradas principalmente en la etapa de preparación de datos.

Preparación de datos

Esta etapa se compondrá de tres elementos principales, anonimización, estandarización y control de calidad de datos:

- **Anonimización:** encaminada a evitar la identificación de las personas involucradas. Es importante que, trabajando con datos sensibles se consideren diferentes técnicas de anonimización, desde métodos más simples, como randomización y enmascarado (Murthy et al., 2019), hasta más específicos para el área de salud, como *k-anonymity* (Olatunji et al., 2022), donde los valores específicos de cada instancia o bien se generalizan sustituyéndolos por otros más generales, o bien se suprimen, eliminándolos para que no se identifique. De este modo, para cada instancia siempre hay al menos otras k – 1 similares.

- **Estandarización:** es el proceso de establecer normas o criterios comunes para que se mantenga la uniformidad y consistencia en un contexto. Se trata de estructura y formato de datos, de modo que tengan un contexto y formateo estable, comprensible y compartible. Contamos con estándares generales a nivel del conjunto de datos , como 'csv', que es el que tomamos en nuestro caso, o más específicos para salud como OMOP-CDM (OHDSI, 2024), empelado para registros médicos o bases de datos clínicas. También existen estándares a nivel de variable, donde el uso de códigos CIE-9 (World Health Organization, 2024) o LOINC (Regenstrief Institute, 2024) se emplea para validar la estandarización, pese a no ser una técnica. Para asegurar la adaptabilidad de la información, es crucial asegurarse de que los datos se encuentran estandarizados de acuerdo con los métodos específicos.
- **Control de calidad de datos:** evaluar y, en la medida de lo posible curar diferentes dimensiones de calidad de datos (Sáez et al., 2012), por ejemplo las dimensiones intrínsecas de, completitud, consistencia, corrección, unicidad y estabilidad. Idealmente, se deberían identificar y evaluar de diferente forma los valores perdidos según su naturaleza (Donders et al., 2006): completamente aleatorios (MCAR, del inglés *Missing Completely at Random*), aleatorios (MAR, del inglés *Missing At Random*) y no aleatorios (MNAR, del inglés *Missing Not At Random*).

  Respecto a la completitud y corrección, de forma técnica, abordamos la existencia de valores perdidos y anómalos (outliers), respectivamente. Para la imputación de valores perdidos, empleamos como solución genérica el uso de Imputaciones Múltiples mediante Ecuaciones Encadenadas (MICE, del inglés *Multivariate Imputation by Chained Equations*). MICE es un método ampliamente utilizado para lidiar con valores perdidos (Stavseth et al., 2019) basado en predecir el valor faltante con un modelo que tome como entrada el resto de características Para la detección de datos anómalos, empleamos percentiles para detectar las anomalías univariantes y un Análisis de Componentes Principales (PCA, del inglés *Principal Component Analysis*) junto con el estadístico Hotelling T^2 (HT2) para las multivariantes. PCA se emplea para proyectar unas variables en un espacio de menor dimensionalidad. Según diversos estudios, el uso conjunto de PCA y HT2 puede ayudar a eliminar los outliers multivariantes de forma muy efectiva (Taskesen, 2023).

  La unicidad se abordaría evaluando la existencia única de las variables asignadas como identificativas en los metadatos. La consistencia mediante la evaluación de reglas de formato, lo que dejaremos para trabajo futuro. La estabilidad se evaluará en la robustez debido a la relación con el aprendizaje. A partir de la fecha de adquisición o la proveniencia de los datos podemos identificar potenciales variaciones en el modelo debido cambios en la fuente o tiempo. Estos cambios podrían evaluarse de forma compleja con herramientas como EHRtemporalVariability (Sáez et al., 2020) o EHRsourceVariability.

Gestión

Una vez pasadas todas las etapas del ciclo de vida e implementado el sistema de IA definitivo, es necesario incluir un **Registro de cambios** y establecer **Protocolos de Acceso**, de modo que tan solo puedan tratar la información aquellas personas formadas y, en cualquier caso, bajo un riguroso registro del tratamiento realizado.

4.2.3. Diversidad, no discriminación y justicia

Una vez realizado el proceso y curado de datos, podemos continuar completando este requisito, cuyo objetivo es permitir la inclusión y diversidad de la IA y aseguramiento de los derechos fundamentales a lo largo de todo el ciclo de vida. Así pues, considerar este requisito en segundo lugar puede ayudar a evitar o cuantificar sesgos injustos antes del modelado.

Preparación de datos

- **Análisis exploratorio sensible:** examinar y visualizar los datos poniendo énfasis en los subgrupos establecidos por las variables sensibles, de tal modo que cualquier mínima irregularidad vinculada a una característica sensible pueda ser detectada, como podría ser la existencia de una diferencia significativa en la presencia de clases por categoría para esa variable o en la correlación con otras variables o diferente. Por ejemplo, podrían emplearse gráficos de barras donde se observe la frecuencia absoluta de cada categoría individual y para cada clase, o evaluar las matrices de correlación obtenidas a partir de cada categoría de la variable sensible.
- **Mitigación de sesgo durante el preprocesado:** aplicar métodos de mitigación de sesgo para minimizar cualquier tipo de asimetría subpoblacional con anterioridad al diseño y entrenamiento del modelo. En concreto, proponemos métodos de sobremuestreo, como la Técnica de sobremuestreo sintético de minorías (SMOTE, del inglés *Synthetic Minority Oversampling Technique*)(Elreedy & Atiya, 2019; Rančić et al., 2021), donde se modifica el conjunto de datos añadiendo nuevas instancias basadas en las ya existentes. Destacar que este método es dependiente del número de datos y el desbalanceo existente, para lo cual podrían emplearse métodos de submuestreo, donde se eliminen instancias del grupo más poblado, o métodos menos potentes de remuestreo simple, donde únicamente se repiten las muestras.

Desarrollo del modelo

- **Mitigación de sesgo durante el procesado:** emplear métodos de mitigación de sesgo para minimizar cualquier tipo de asimetría subpoblacional durante la fase de entrenamiento y diseño del modelo. Un enfoque práctico podría ser realizar una reponderación (*reweighting*) para asignar unos pesos a las categorías acorde con su presencia en el conjunto total de datos (Krasanakis et al., 2018). Mencionar que para la reponderación se emplea la frecuencia de las categorías. Sin embargo, podrían emplearse multitud de funciones de coste que establezcan jerarquías para los diferentes tipos de errores. Esta aproximación es muy frecuente en redes neuronales, donde se ajustan los pesos de la red para penalizar de manera diferenciada los errores según su impacto en la equidad del modelo.
- **Rendimiento por variables sensibles:** analizar de forma aislada el rendimiento asociado a cada subgrupo dentro de las diferentes variables sensibles con el fin de determinar asimetrías subpoblaciones. En caso de encontrarse diferencias, se deberían reevaluar los métodos de mitigación de sesgo y, en cualquier caso, suspender la implementación del modelo para evitar predicciones injustas.

Implementación y uso

- **Mitigación de sesgo durante el posprocesado:** emplear métodos de mitigación de sesgo para minimizar cualquier tipo de asimetría subpoblacional durante la implementación del modelo, de modo que se lidie correctamente con los posibles nuevos casos. Proponemos un enfoque de clasificación *Reject Option Based* (Kamiran et al., 2018) donde, en aquellas muestras pertenecientes a una categoría discriminada, en caso de no poder atribuirse su clasificación de forma rotunda, se le asignará de forma conservadora la clase positiva.
- **Monitorización de Justicia:** incluir herramientas para poder detectar de forma continua cualquier tipo de variación en el nivel de justicia alcanzado, de modo que siempre se mantenga un nivel mínimo. Se debería de, para las variables sensibles, evaluar tanto el rendimiento del modelo como las propias distribuciones de los datos. Para ello, podría emplearse la herramienta EHRtemporalVariability (Sáez et al., 2020) o métodos basados en ella.

De un modo transversal aunque desde el mismo punto de vista de la Diversidad, no discriminación y justicia, nos hemos asegurado de que en aquellos gráficos donde se requiera un mínimo de interpretación basada en los colores, se empleen paletas aptas para daltónicos, como es el caso de "viridis" o configuraciones "colorblind".

4.2.4. Transparencia

Seguidamente, podemos continuar completando este requisito, cuyo objetivo es permitir la inclusión y diversidad de la IA a lo largo de todo el ciclo de vida. Así pues, considerar este requisito en 2º lugar puede ayudar a evitar o cuantificar sesgos injustos.

Preparación de datos

- **Procedencia de los datos:** extrae la información incluida en los metadatos referente al linaje de los datos e intentaremos emplearla para adecuar y comprender el contexto a nivel técnico.
- **Análisis exploratorio general:** examinar y visualizar los datos con el fin de identificar cualquier correlación entre variables y con la clase a predecir. Se detectarán aquellas características relevantes y las que, por el contrario, son redundantes. Puede realizarse para las numéricas mediante una matriz de correlaciones o un gráfico bivariante y para las categóricas mediante diagramas de barras donde se observe la proporción de cada clase por categoría

Desarrollo del modelo

- **Describir los pasos de diseño:** documentar y explicar el uso de los modelos a nivel técnico, tanto el porqué de su elección como la justificación de adopción de sus parámetros. También deben incluirse los métodos de entrenamiento y validación escogidos. Para ello, puede emplearse por ejemplo una plantilla TRIPOD (Collins et al., 2024).
- **Gráficos de explicabilidad:** emplear gráficos que, a nivel de entrenamiento, sean útiles para comprender las decisiones del modelo, como el uso de matrices de confusión, donde se enfrentan las etiquetas reales y predichas, o curvas características operativas del receptor (ROC, del inglés *Receiver Operating Characteristic*), donde se muestra la relación de

especificidad y sensibilidad del modelo para distintos umbrales. También pueden emplearse gráficos de barras para observar la importancia de las variables en el entrenamiento. De permitirlo y ser útil, también podríamos incluir gráficos específicos, como visualizaciones del funcionamiento de un árbol en Random Forest.

Implementación y uso

- **Gráficos de explicabilidad:** tras finalizar el entrenamiento. Para completar la compresión del modelo y complementar los resultados, empleamos métodos agnósticos que permitan explicar el funcionamiento de los modelos, que ayuden a entender las decisiones, para facilitar la comprensión de personal ajeno al diseño. Empleamos pues las explicaciones aditivas Shapley (SHAP, del inglés *SHapley Additive exPlanations*), las cuales permiten explicar las predicciones individuales basándose en unos valores Shapley (Molnar, 2021). Estos valores Shapley se interpretan como la contribución de cada característica a una predicción en particular, asignando una importancia a cada característica al comparar su efecto con todas las posibles combinaciones de características. Esto ayuda a entender cómo cada característica influye en el resultado, lo que es útil para identificar patrones y posibles sesgos en las decisiones del modelo.

Gestión

Aprovechamos para incluir en la etapa de gestión una advertencia o ***disclaimer*** donde se informe acerca de las características del modelo y las posibles limitaciones que pueda tener. La plantilla propuesta es la siguiente:

DISCLAIMER FOR TRUSTWORTHY AI MODEL [Model Name]

This trustworthy AI model, [Model Name], has been meticulously developed for [intended use cases]. It underwent extensive training on [describe the data] and consistently demonstrates a high level of accuracy, with performance metrics indicating [mention performance metrics] when assessed in [describe the context or domain]. The model has undergone rigorous testing for fairness and robustness, although it's important to acknowledge [include any limitations regarding fairness, bias, and robustness].

[Model Name] employs inherently [transparent/ not transparent] algorithms to the end-user, and we employ [describe any interpretability tools or methods used] to ensure transparency and interpretability. Data privacy and security are paramount, and we uphold strict measures [describe mesures] to safeguard user data in compliance with regulations.

While [Model Name] has been carefully designed to minimize biases and ethical concerns, users are urged not to rely solely on it for [critical decisions/ any sensitive use cases]. It is imperative to exercise caution when interpreting the model's predictions and, when appropriate, consult with a qualified human expert for review.

To maintain your trust in [Model Name], we are dedicated to routine updates and maintenance, ensuring it remains aligned with the latest data and best practices. Should you have any inquiries, concerns, or encounter any issues, please do not hesitate to reach out to us at [contact information].

Please take note that [Company/Developer Name] cannot be held liable for any harm or damage resulting from the use of [Model Name] outside of its designated use cases and recommendations.

[Include any additional specific disclaimers related to the model].

Es importante también, una vez concluida la implementación del modelo, **Documentar los registros** de los resultados de los modelos predictivos durante su uso. Esto es especialmente importante para monitorizar el funcionamiento del modelo y poder registrar cualquier tipo de fallo. Del mismo modo, se debe ofrecer una vía para **Compartir incidentes**, ya que en vista de casos imprevistos es posible que sucedan incidentes no contemplados potencialmente perjudiciales para los usuarios finales o terceros.

4.2.5. Robustez técnica y seguridad

Por último, el requisito de robustez técnica y seguridad tiene como objetivo evitar o minimizar cualquier daño involuntario o inesperado. Para ello, se emplearán todos aquellos métodos que mejoren el rendimiento del modelo y que permitan cuantificar de forma robusta su comportamiento.

Preparación de datos

- **Selección de características o Reducción de la dimensionalidad:** para lidiar con la maldición de dimensionalidad (Karanam, 2021) es de gran utilidad emplear métodos para encontrar aquellas características más relevantes, como el uso de selección de características mediante *filters* o *wrappers*, o mediante métodos de extracción de características como PCA (Christopher M. Bishop, 2006; R. Duda et al., 2001), donde proyectamos las características existentes es un espacio de mayor variabilidad y menor dimensionalidad. De no realizarse esta fase, el volumen de datos podría acarrear problemas de robustez o eficiencia.
- **Balanceo de clases:** proponemos métodos de sobremuestreo o submuestreo como SMOTE para lidiar con las diferencias de frecuencia absoluta entre las distintas clases. De existir un desbalanceo entre clases y no corregirlo, podríamos encontrar un sesgo en las predicciones del modelo hacia la clase sobrerrepresentada.

Desarrollo del modelo

- **Optimización de hiperparámetros:** habitualmente, el desarrollo de modelos de IA, es común realizar una búsqueda de parámetros para encontrar aquellas que obtienen el mejor resultado para una función de coste dada. Pese a que es frecuente realizar la optimización manualmente, en la actualidad se disponen de técnicas automáticas ampliamente utilizadas como *Gridsearch* (Belete & D H, 2021), que con un modelo, una red de parámetros y una métrica a maximizar, encuentra la mejor combinatoria. Por ejemplo, para Random Forest podríamos buscar el número óptimo de árboles de decisión o la mejor profundidad para cada uno de ellos. En trabajo futuro podrían emplearse técnicas de *Meta-learning* (Vilalta et al., 2010) donde, a diferencia del aprendizaje base donde se acumula experiencia para una tarea de aprendizaje específica, el *Meta-learning* se centra en buscar patrones y estrategias aplicables a diferentes contextos.
- **Cuantificación de la incertidumbre:** es imprescindible que el usuario conozca tanto la exactitud como la precisión del modelo, es decir, cómo de bueno es el rendimiento y cómo de variable es este rendimiento. Para ofrecer la cuantificación de la incertidumbre en la fase de entrenamiento obtenemos numéricamente los intervalos de confianza de las métricas de entrenamiento. Puede emplearse el método *Bootstrap* para entrenar diferentes modelos basados en un subconjunto de los datos de entrenamiento con reemplazo a partir de los cuales

estudiar la variabilidad de las métricas (Endo et al., 2015). Mencionar que, de ser significativo el tamaño del conjunto de datos, podría emplearse un método de validación cruzada (CV, del inglés *Cross Validation*).

- **Evaluación bien formada:** realizar una evaluación mediante métodos representativos como validación cruzada (Bradshaw et al., 2023), el cual se basa en calcular el promedio de modelos entrenados y evaluados con subconjuntos dentro de los datos de entrenamiento, y empleando métricas robustas.
- **Curva de calibración:** para evaluar la fiabilidad de las predicciones. La curva compara las probabilidades predichas por el modelo con la frecuencia real para diferentes umbrales de probabilidad. Para formar la curva se dividen las predicciones en intervalos, se calcula la proporción de resultados verdaderos de cada uno y se trazar la probabilidad promedio predicha frente a la frecuencia real, se visualiza la calibración del modelo. Predicciones perfectamente calibradas tendrían una recta y=x, las desviaciones se deben sobre o subestimaciones.

Implementación y uso

- **Cuantificación de la incertidumbre:** para ofrecer cuantificación de incertidumbre en la fase de predicción empleamos los intervalos de confianza diseñados en el entrenamiento, junto con el valor numérico de la incertidumbre existente en las predicciones. Puede ser útil predecir un dato con los múltiples modelos entrenados por *Bootstrap* para ver su coherencia (Método Bayesiano) (Abdar et al., 2021; Endo et al., 2015) o perturbar un dato múltiples veces para observar como hace frente el modelo a la variabilidad (Método de Aumentado, o *data augmentation* en inglés) (Abdar et al., 2021).
- **Monitorización de *dataset shifts* en el conjunto de datos:** evaluar la posible variabilidad de los datos a lo largo del tiempo o según nuevos contextos. Las distribuciones podrían verse afectadas por la temporalidad o cambio de contexto de aplicación, de modo que las predicciones y modelos también disminuyan su rendimiento en igual medida. Existen tres tipos de *dataset shifts* (Huyen, 2022): *concept shift*, cuando cambia la relación entre la entrada y la salida*; covariate shift*, cuando las distribuciones de las características cambian manteniendo la relación entre entradas y salidas constante; y *prior probability shift*, cuando las distribuciones de las clases cambian manteniendo la relación entre entradas y salidas constante. La variación en las distribuciones podría realizarse empleando herramientas como EHRtemporalVariability o EHRsourceVariability (Sáez et al., 2020)
- **Reentrenamiento:** considerar la actualización del entrenamiento con nuevas versiones de los datos en caso de que se produzca algún tipo de variabilidad en el contexto, como temporal o geográfica. Profundamente relacionado con el concepto de *dataset shift*, ya que comprende la acción a realizar en caso de que, tras la monitorización del conjunto de datos, se identifiquen variaciones. El reentrenamiento podría ser completo, olvidando lo aprendido previamente, o empleando técnicas de *continual learning*, donde el modelo adapta sus conocimientos al recibir nueva información.

- **Seguridad general:** evitar que el modelo realice predicciones imprecisas, impedir la predicción en caso de existir inconsistencias. En primer lugar, la seguridad puede abordarse desde un punto de vista previo al desarrollo del modelo, donde comparando la distribución de un nuevo dato con la de los datos de entrenamiento podemos determinar si es un *outlier* y por tanto si es susceptible de predecirse de forma más inexacta. En segundo lugar, a la hora de realizar predicciones, podemos observar la distribución de probabilidad de la predicción dato. Si el índice de saturación se encuentra dentro de un intervalo de confianza del 95% de la predicción, podemos afirmar que no existe consistencia suficiente para clasificar el dato.

4.3. Evaluación en conjunto de datos "Diabetes"

A continuación, mostramos los resultados de evaluar los pipelines propuestos para el "conjunto Diabetes".

4.3.1. Metadatos

El "conjunto Diabetes" cuenta con una gran variedad y cantidad de datos, por lo que, resulta muy interesante para la aplicación en concreto, si bien hemos simplificado algunas de las variables para permanecer ajustados al objetivo de este trabajo. En primer lugar, eliminamos todas aquellas características referentes a la variación de un medicamento en particular, ya que es un volumen considerable del conjunto de datos y podemos asumir que, con la variable 'change', que indica si hay o no variación en alguna de las medicinas, se puede conservar la mayor parte de la información.

En segundo lugar, la base de datos está compuesta por encuentros. Por tanto, tenemos tantos encuentros como número de datos y un paciente asociado a cada encuentro, permitiendo que aparezca más de una vez el ID de un paciente. Como nuestra variable a predecir es el reingreso <30 días, asumiendo que el tiempo en el hospital será una variable muy relacionada con él, conservamos, de entre las instancias con mismo ID de paciente, la que presenta el máximo tiempo en el hospital.

Seguidamente, hay variables con un volumen de categorías muy grande, permitiendo que haya valores existentes en porciones incluso <1% del dataset. Por tanto, realizamos una simplificación de las categorías de las variables 'race', 'discharge_disposition_id', 'admission_source_id' y 'admission_type_id' agrupando categorías poco representadas en otras menos específicas. Además, realizamos la traducción de los IDs a texto para las altas y las admisiones. Por último, convertimos la variable 'age', agrupada por intervalos, en valores numéricos según la mediana del intervalo.

En la *Tabla 1* podemos observar los Metadatos resultantes que se introducirán al comienzo del pipeline por el script de Privacidad y gobernanza de datos. Este dataset resulta considerablemente completo ya que permite cumplimentar todos los campos propuestos en la metodología general, contando con variables identificativas, sensibles, proveniencia de los datos, y clase positiva con la que binarizar el problema multiclase.

Tabla 1. Metadatos iniciales para el "conjunto Diabetes".

METADATOS PARA EL "CONJUNTO DIABETES"	
Conjunto de datos	'dataset_diabetes_simplified.csv'
Salida	'readmitted'
Clase positiva	'<30'
Características identificativas	'encounter_id' y 'patient_nbr'
Características sensibles	'race' y 'gender
Característica para balancear	'race'
Proveniencia de los datos	["Una base de datos de Health Facts que representa 10 años (1999-2008) de atención clínica en 130 hospitales de Estados Unidos." , 'admission_type']
Fecha de adquisición	["Vacía", " "]
Tipos de características	▪ race: categórica ▪ gender: categórica ▪ age: numérica ▪ weight: categórica ▪ admission_type: categórica ▪ discharge_disposition: categórica ▪ admission_source: categórica ▪ time_in_hospital: numérica ▪ payer_code: categórica ▪ medical_specialty: categórica ▪ num_lab_procedures: numérica ▪ num_procedures: numérica ▪ number_outpatient: numérica ▪ number_emergency: numérica ▪ number_inpatient: numérica ▪ diag_1: categórica ▪ diag_2: categórica ▪ diag_3: categórica ▪ number_diagnoses: numérica ▪ change: categórica ▪ diabetesMed: categorical

4.3.2. Privacidad y gobernanza de datos

A partir del conjunto de datos proporcionado, en primer lugar, eliminamos las variables identificativas ('encounter_id' y 'patient_nbr'), ya que, a pesar de estar anonimizadas (**Anonimizado**), a priori no otorgan información útil para la predicción.

A continuación, observamos que tenemos variables codificadas según códigos diagnósticos, por tanto, basándonos en su CIE-9 realizamos su estandarización ajustándonos a su categoría médica (**Estandarizado**). El código CIE-9 hace referencia a una patología muy específica, por lo que, para reducir la variabilidad de categorías, empleamos la agrupación por intervalos propuesta en la bibliografía (World Health Organization, 2024).

Por último, realizamos un **control de la calidad de los datos**. Estos satisfacen la dimensión de Unicidad, al no existir datos duplicados, y tienen una baja actualidad, ya que los datos tienen en torno a 2 décadas, si bien para el propósito actual no es una limitación. Entrando en dimensiones más complejas como la Corrección o Completitud, realizamos el correspondiente análisis de outliers y valores perdidos, respectivamente. En primer lugar, obviamos todas las características con una cantidad de valores perdidos superior a un umbral, que en este caso situamos en 40% (*Figura 9*).

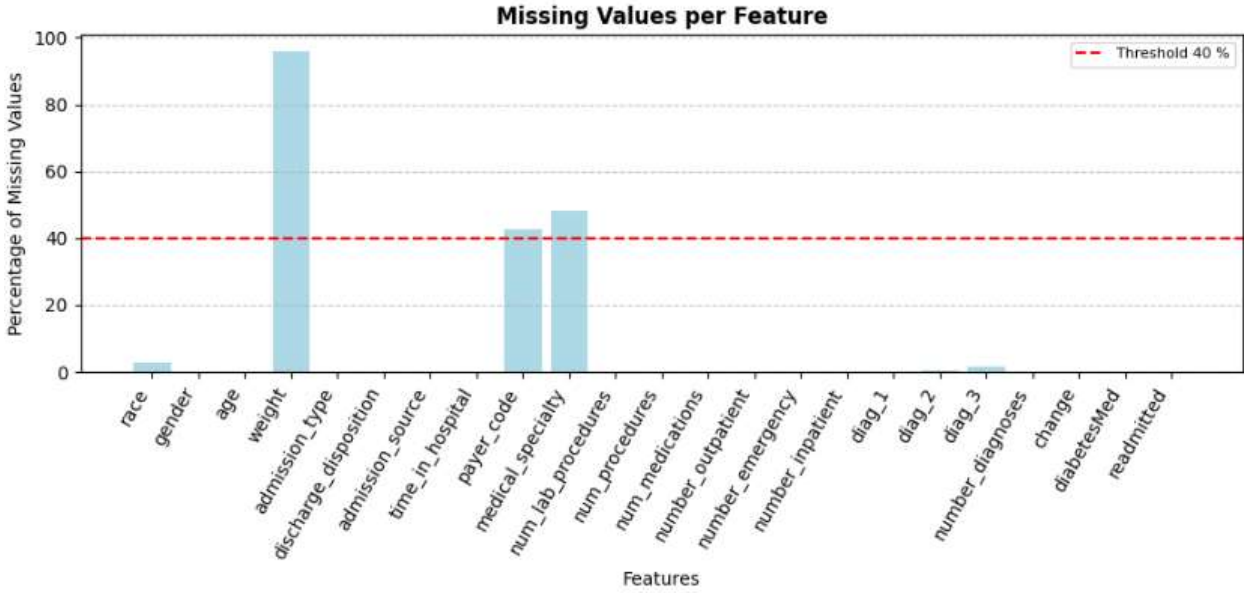

Figura 9. Gráfico de barras de valores perdidos con umbral de eliminación en 40% perdidos en el "Conjunto Diabetes".

Con el fin de sesgar lo menos posible todo el flujo de trabajo, realizamos en este punto una partición del conjunto total de datos en un subconjunto de entrenamiento del 90% y 10% de test. Todas las posteriores acciones se realizarán con los datos de entrenamiento

A continuación, analizamos los outliers univariantes de las características numéricas estableciendo un percentil 95%, de forma que todos los elementos situados en las colas de la distribución susceptibles de ser outliers se sitúen como valores perdidos. El total de celdas con outliers detectadas es de 9950, lo cual supone un 1.75% del total de celdas numéricas. En la *Figura 10* podemos ver los box-plots resultantes.

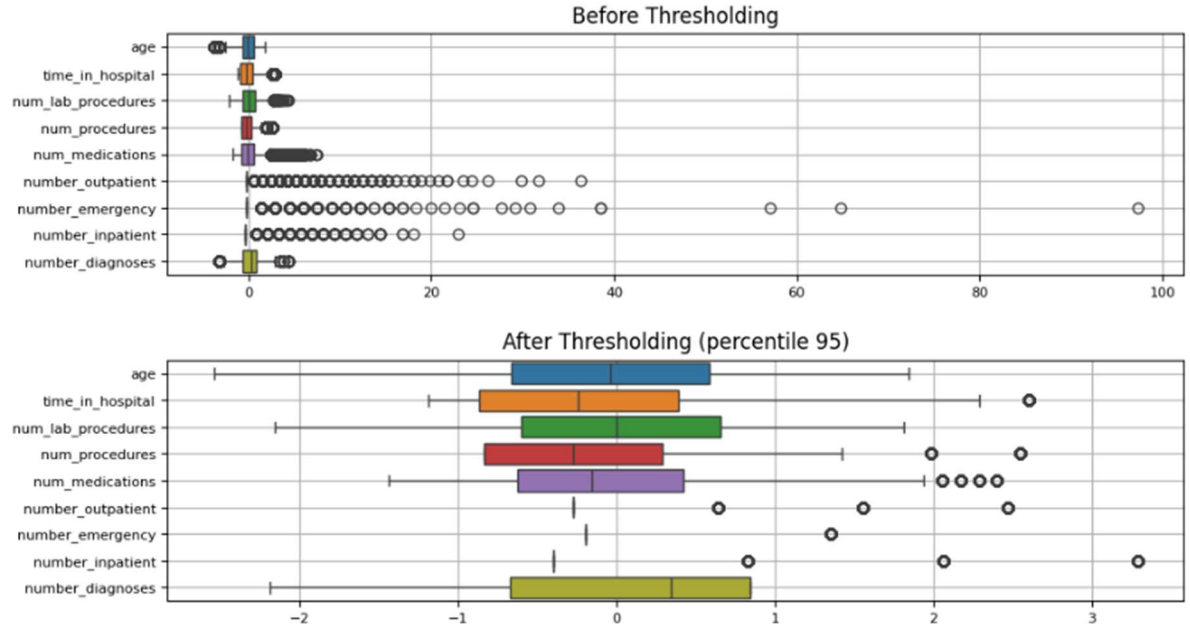

Figura 10. Box-plot de las características numéricas normalizas antes y después de umbralizar en el "Conjunto Diabetes".

El anterior gráfico también aporta información acerca de la relevancia que pueden tener las variables. Aquellas con valores constantes o con baja variabilidad pueden no tener relevancia en las decisiones del modelo. Es este el caso de las características 'number_outpatient', 'number_emergency' y 'number_inpatient' que tienen para la gran mayoría de instancias valor 0. Podemos ver la distribución por clase de algunas de estas variables en la *Figura 11.*

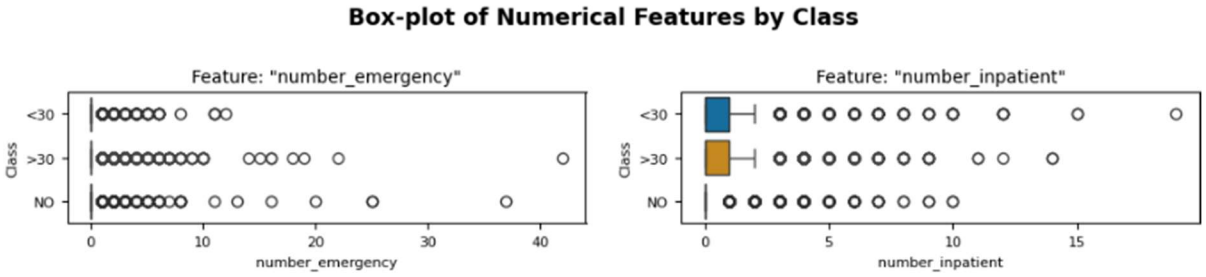

Figura 11. Box plot de características numéricas por clase

Para abordar los datos anómalos categóricos, graficamos las frecuencias absolutas de cada categoría. Determinamos que aquella categoría presente en un número menor a un umbral de 0.1% del total es susceptible de ser eliminada. En la *Figura 12* vemos como la variable 'gender' tiene únicamente 3 instancias con 'Unknown/Invalid', por lo que, al estar por debajo del umbral eliminamos las filas con esa categoría.

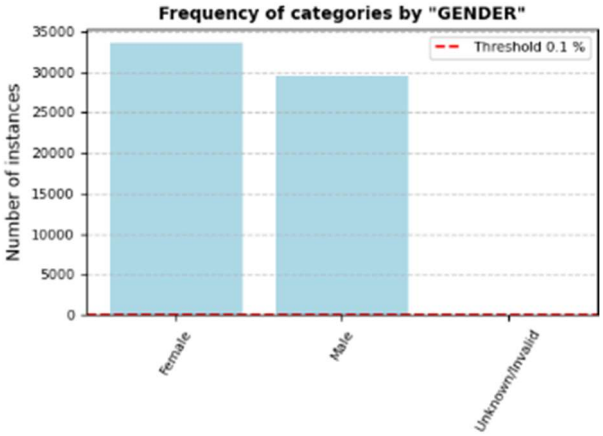

Figura 12. *Gráfico de barras de las frecuencias absolutas de cada categoría en la variable 'gender' en el "Conjunto Diabetes".*

Seguidamente, solventamos los anteriores datos perdidos junto con los encontrados inicialmente, prediciendo sus valores mediante el método MICE. En este caso, empleamos un método de regresión lineal para las variables numéricas y un K vecinos próximos (KNN, del inglés *K-Nearest Neighbor*) para las categóricas. Debido a que al inicio del documento no se ha diferenciado entre variables numéricas y discretas, ni categóricas nominales u ordinales, puede ser que los métodos de imputación no sean los óptimos para cada tipo de variable. Sin embargo, no realizamos modificaciones para no afectar a la generalización. Es importante también destacar en este punto la imputación de los posibles valores perdidos en el conjunto de test a partir de los modelos entrenados con los datos de entrenamiento.

Para finalizar, realizamos un análisis multivariante donde por medio de PCA proyectas las características en un nuevo espacio dimensional y eliminamos aquellos datos que presentan un nivel de significancia α=0.01, resultando así un total de 809 outliers multivariantes. En la *Figura 13* observamos la proyección de los puntos con las tres primeras componentes principales.

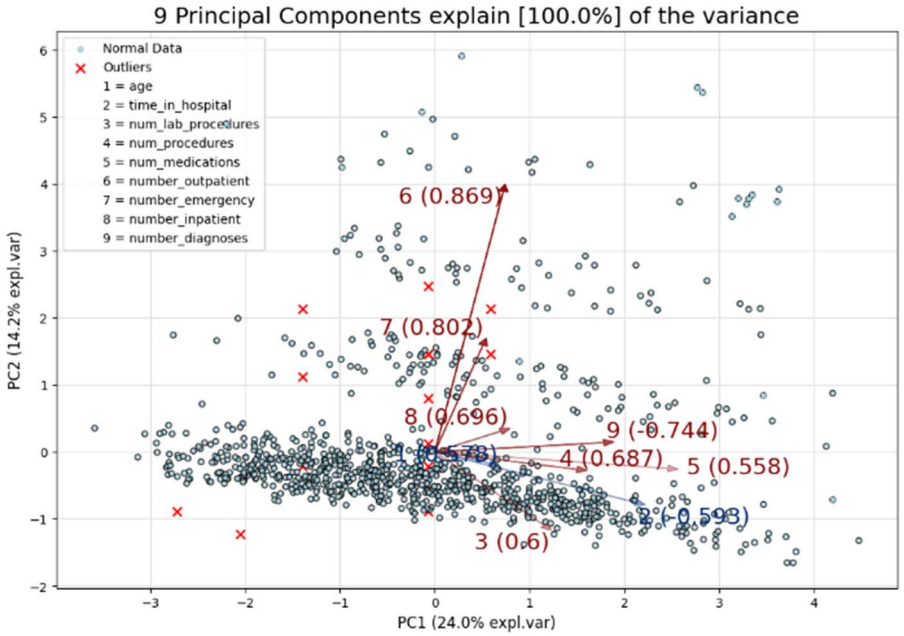

Figura 13*. Gráfico de dispersión 3D de las tres primeras componentes principales en el "Conjunto Diabetes".*

Además, aprovechando que contamos con todos los datos curados y carecemos de la variable "acquisition_date" vamos a simular la fecha de adquisición realizando un PCA mixto o FAMD (del inglés *Factor Analysis of Mixed Data*). De este modo, seleccionamos la componente de variabilidad principal y segmentamos sus valores con 10 intervalos que representan los 10 años de adquisición de los datos. En la *Figura 14* vemos un subconjunto de 10000 datos donde a cada dato le asignamos un año según al intervalo al que pertenece.

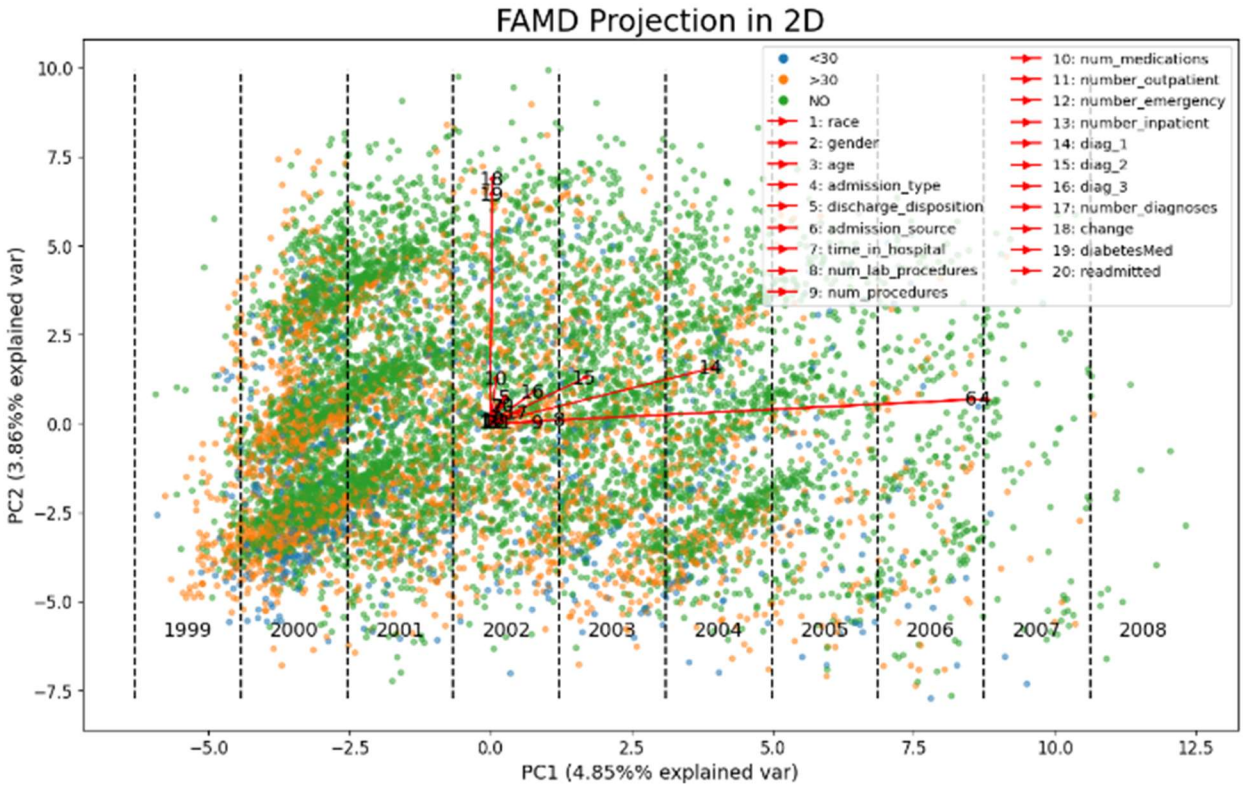

Figura 14. *Proyección FAMD las 2 componentes principales de un subconjunto del "Conjunto Diabetes"*

4.3.3. Diversidad, no discriminación y justicia

Tras realizar el tratamiento de datos, podemos comenzar a estudiar los sesgos potenciales vinculados a las variables sensibles. Comenzamos realizando un **análisis exploratorio sensible**, para lo cual empleamos un gráfico de barras que nos permite ver las distribuciones totales por categoría y clase en cada variable sensible, tal y como se aprecia en la *Figura 15.* En ella, se observa un claro desbalanceo en la presencia de las diferentes razas y cierto desequilibrio en el sexo.

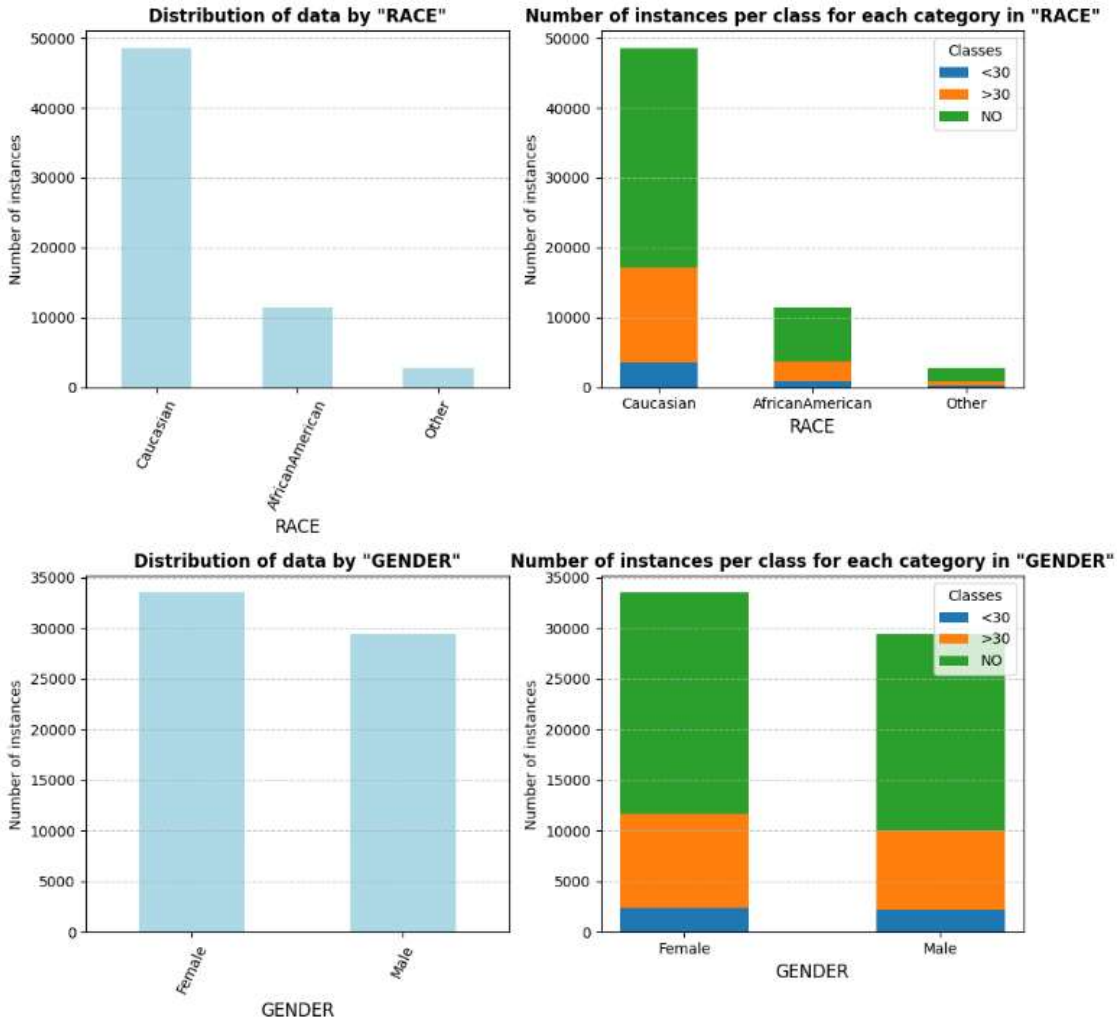

Figura 15. Distribución de datos por categoría y clase para cada variable sensible en el "Conjunto Diabetes".

Seguidamente, con los valores perdidos detectados en la etapa de Privacidad y Gobernanza de datos, podemos graficar su distribución según las categorías sensibles. En la *Figura 16* se aprecia como, a nivel general, las razas 'Other y 'AfricanAmerican' siempre presentan al menos un valor perdido por instancia.

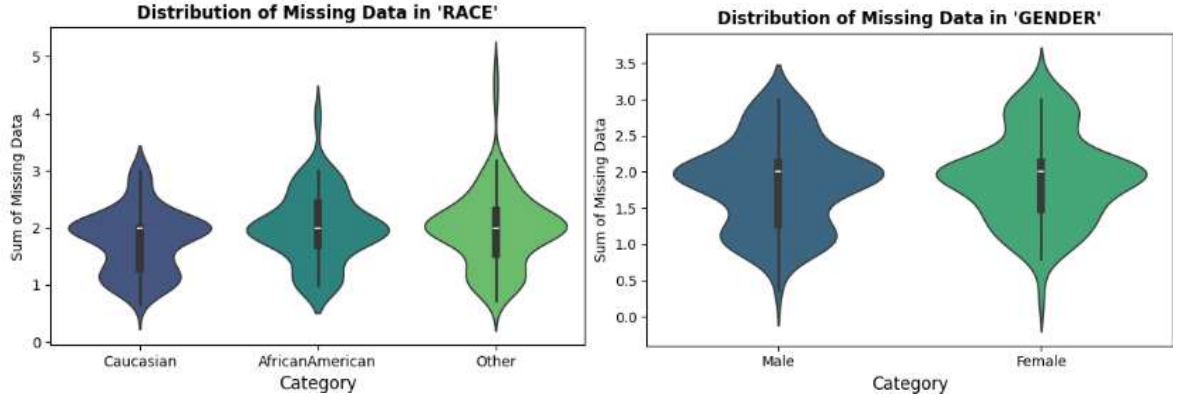

Figura 16. Distribución de valores perdidos por categoría según las variables sensibles en el "Conjunto Diabetes".

Para terminar el análisis exploratorio, también proponemos observar las diferencias de correlación de las variables numéricas para las distintas categorías sensibles. En la *Figura 17* identificamos como, en este caso, las personas con raza 'AfricanAmerican' cuentan con una correlación 'num_procedures-age' 0.09 puntos mayor a las que tienen raza 'Caucasian'.

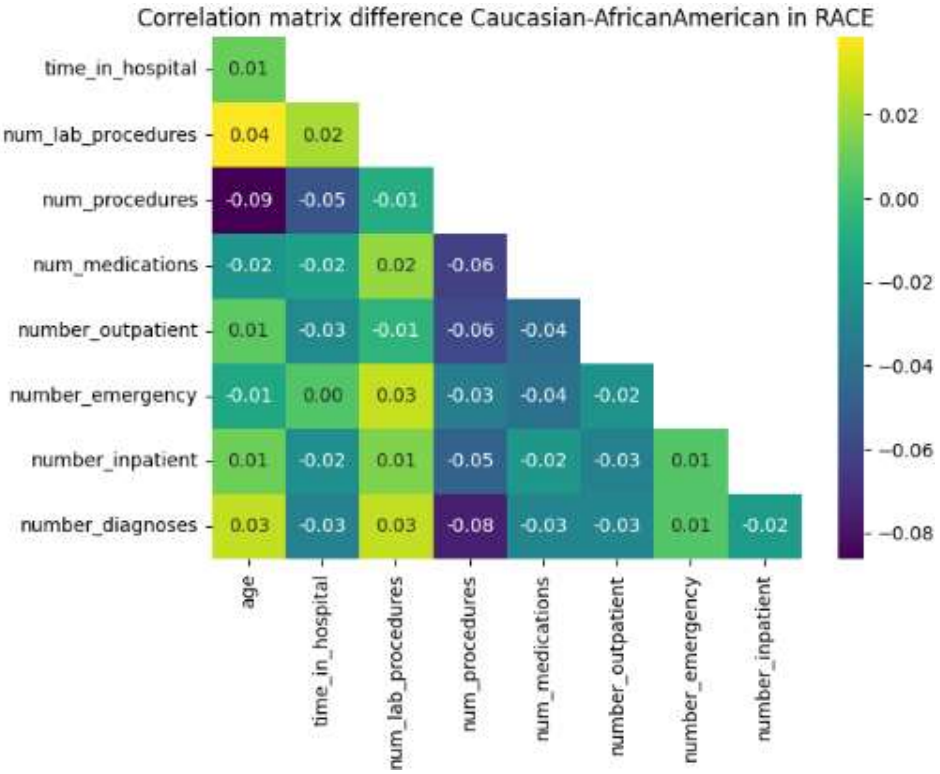

Figura 17. *Mapa de calor de la diferencia de las matrices de confusión obtenidas con las categorías 'Caucasian' y 'AfricanAmerican' en la variable sensible 'race' en el "Conjunto Diabetes".*

A continuación para eliminar el posible sesgo y comparar los resultados obtenidos sin tratar el sesgo, realizamos una comparación de 3 caminos posibles: obviar la existencia de sesgos y trabajar con los datos de entrenamiento y un modelo convencional, aplicar **técnicas de mitigación de sesgo en el preprocesado** (como el sobremuestreo mediante SMOTE) y **técnicas de mitigación de sesgo durante el procesamiento** (como la reponderación o *reweighting*).

En la *Figura 18* podemos ver el resultado de entrenar 3 modelos distintos siguiendo cada una de las rutas explicadas. De ella extraemos que curvas ROC muy similares, lo cual puede indicar que no existe sesgo por la raza en estos datos, ya que la variable raza podría seguir siendo importante, por ejemplo viendo su importancia posteriormente

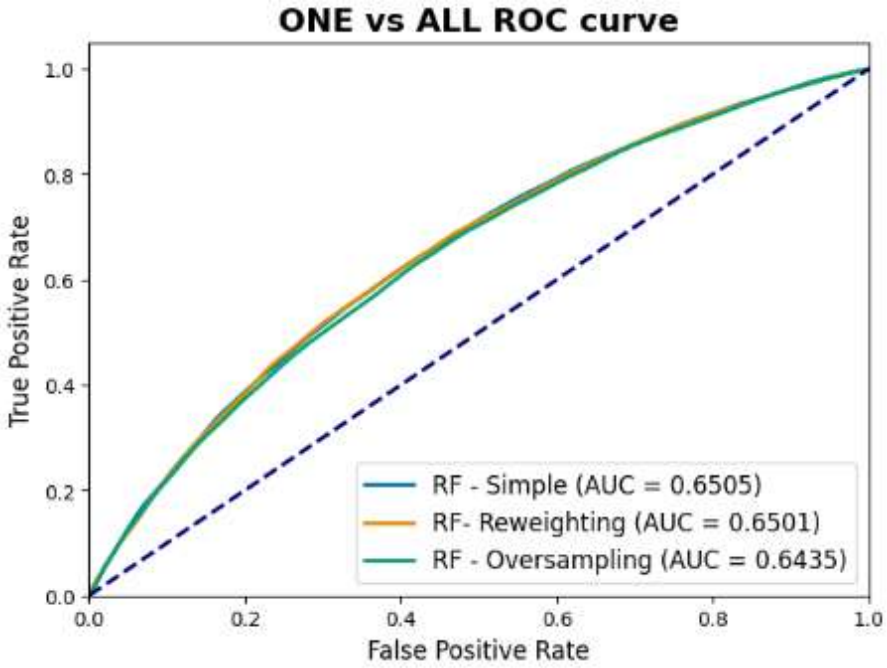

Figura 18. *Comparación curvas ROC para un modelo simple, con reponderación y con sobremuestreo en el "Conjunto Diabetes".*

Si observamos como predice un modelo entrenado con todas las razas para cada raza, observamos que es no existen diferencias previo y tras la mitigación de sesgo con sobremuestreo (*Figura 19*).

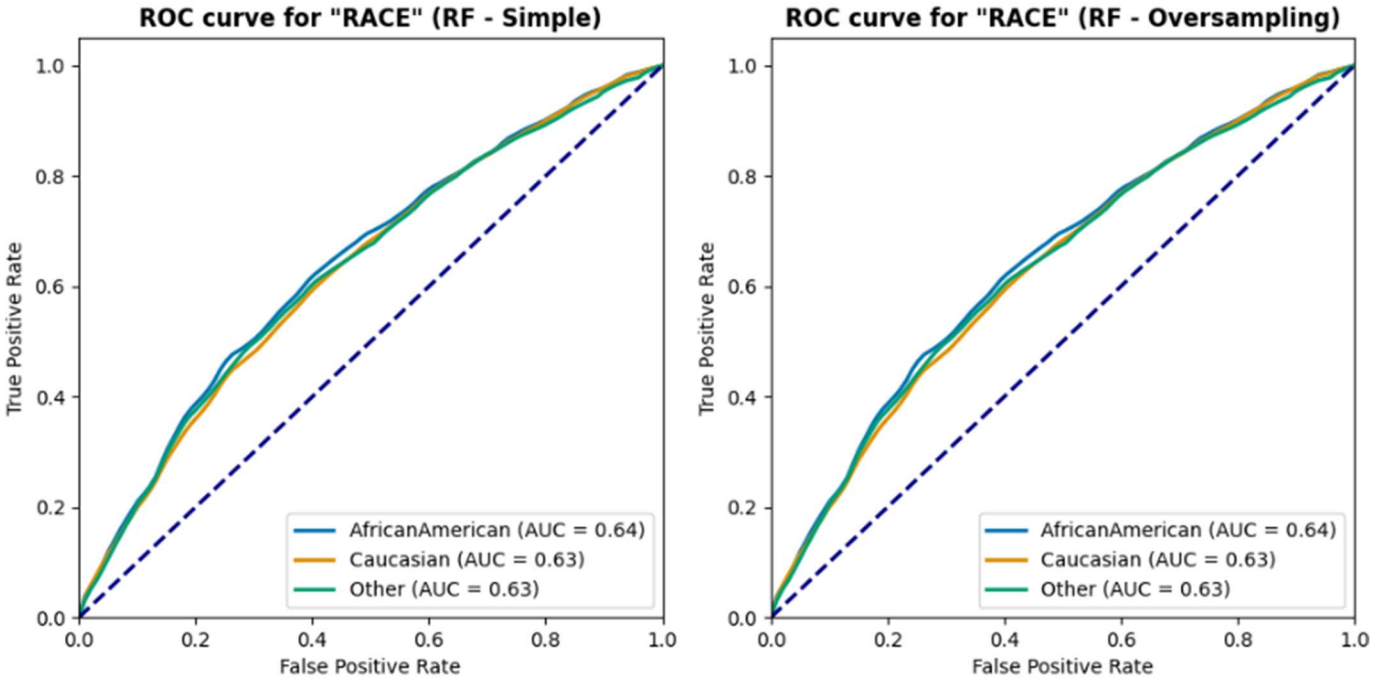

Figura 19. Predicciones de datos con distinta raza de modelo entrenado normal y con sobremuestreo.

A continuación, probamos a entrenar distintos modelos con datos únicamente pertenecientes a cada categoría sensible, obteniendo un modelo para cada una de las razas 'Caucasian', 'AfricanAmerican' y 'Others' así como uno para cada sexo 'Male' y 'Female'.

De este modo, en la *Figura 20* podemos analizar cómo es la importancia de las variables para cada modelo:

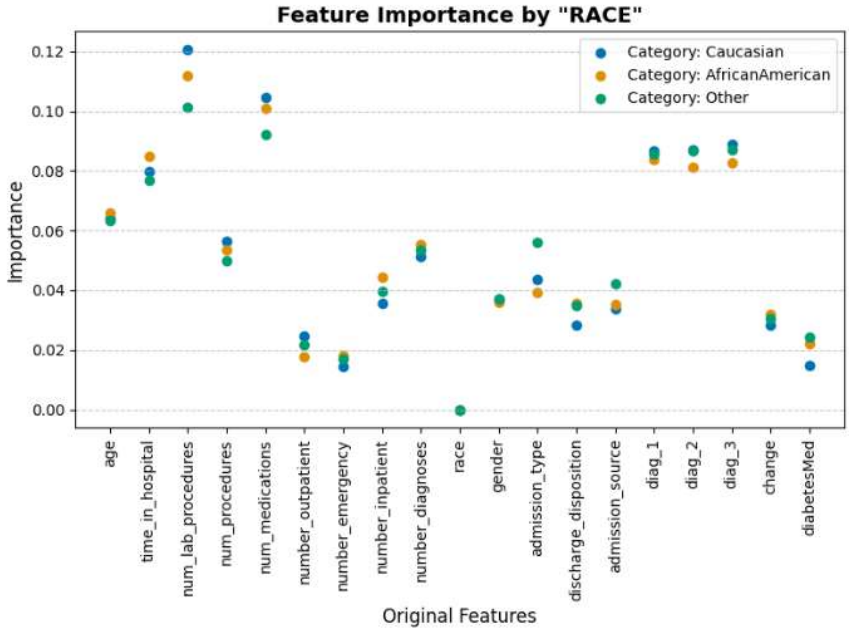

Figura 20. *Importancia en la predicción de cada variable según la raza en el "Conjunto Diabetes".*

Por último, también planteamos un método de mitigación de sesgo en el posprocesado, donde, basándonos en un enfoque de clasificación *Reject Option Based*, etiquetamos con la clase positiva todas las predicciones que se sitúen cerca de un umbral de decisión en un subgrupo discriminado. Para la ejemplificación, se ha supuesto la variable menos representada en la característica a balancear 'race'. Sin embargo, esta aplicación debería realizarse en aquellos contextos donde se haya demostrado con los métodos descritos anteriormente que existe un sesgo real.

4.3.4. Transparencia

Una vez tenemos controlada la calidad y el sesgo de los datos, podemos comenzar a entender cómo éste se comporta. Así pues, empleamos el script de Transparencia con el fin de alcanzar la máxima comprensión tanto de los datos como del modelo y sus predicciones. En este punto cabe mencionar de nuevo la distinción entre explicabilidad e interpretabilidad. Idealmente, intentaremos realzar ambas. Sin embargo, así como el inicio del script irá encaminado a otorgar principalmente interpretabilidad, los gráficos finales de soporte a la predicción tienen un fin principalmente explicativo, de modo que el usuario final pueda entender de modo sencillo qué está influyendo en la decisión del modelo (European Comission, 2020). Destacamos que el script cuenta con la posibilidad de binarizar el problema, permitiendo por tanto acceder a información no accesible para problemas multiclase.

En primer lugar, observamos la **proveniencia de los datos** accediendo a la variable 'data_provenance' incluida en los metadatos con el fin de buscar algún tipo de información relevante acerca de la procedencia de los datos. Conocer el contexto de cada instancia y del conjunto en general puede ayudar a controlar y evaluar la calidad de los datos En nuestro caso, no aporta conocimiento significativo aunque ayuda a contextualizar la aplicación.

Seguidamente, continuamos observando de nuevo la distribución de los datos, buscando posibles interrelaciones y redundancias. El **análisis exploratorio** realizado consta de gráficos independientes para las variables numéricas y categóricas. Para las numéricas, en la *Figura 21* observamos el graficado mediante un mapa de calor de la matriz de correlación numérica.

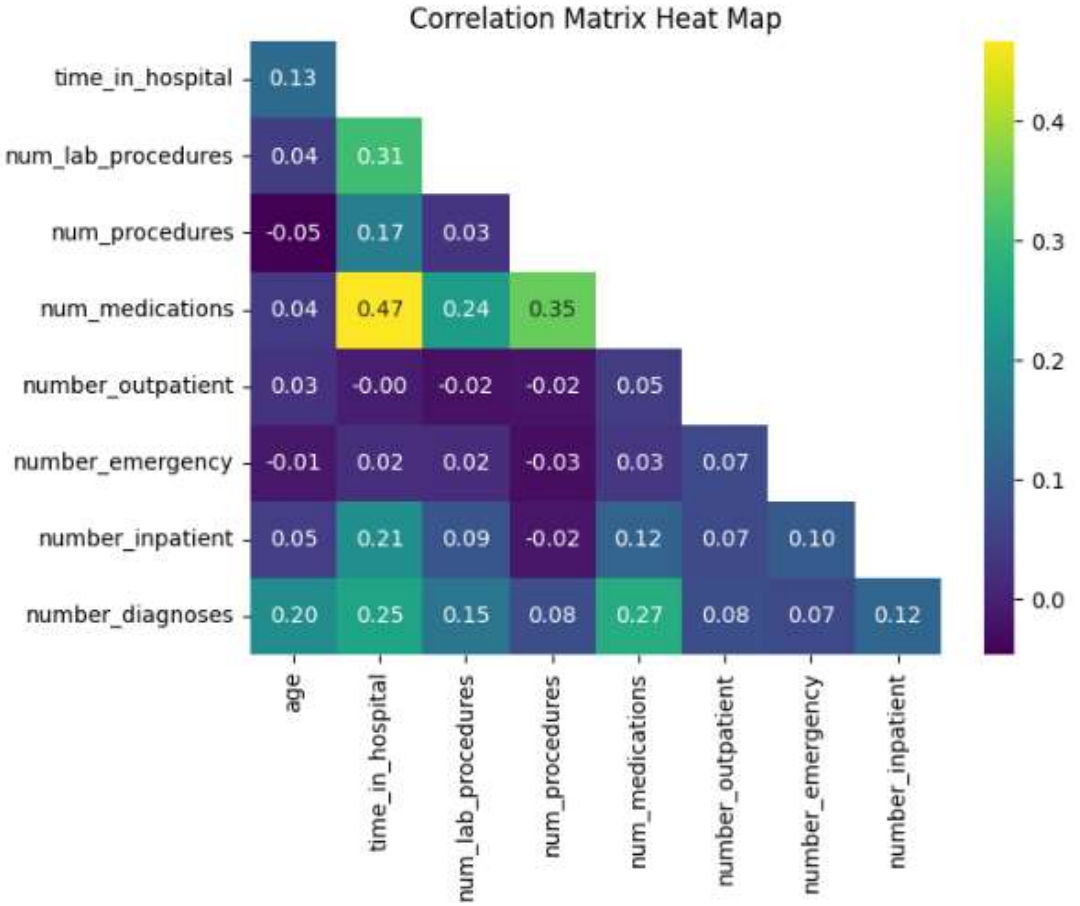

Figura 21. *Mapa de calor de la matriz de correlaciones de las variables numéricas en el "Conjunto Diabetes".*

De la anterior figura no se puede afirmar que exista ningún tipo de correlación significativa. Por otro lado, para las variables categóricas, graficamos mediante diagramas de barras apilados la proporción de las clases según cada categoría con el fin de determinar aquellas variables más discriminativas. En la *Figura 22* apreciamos algunos resultados. Por ejemplo, en la variable 'diag_1', si el paciente tiene como diagnóstico principal diabetes, este tendrá una menor probabilidad de ingreso '<30', por lo que esta podría ser buena categoría para descartar esta clase.

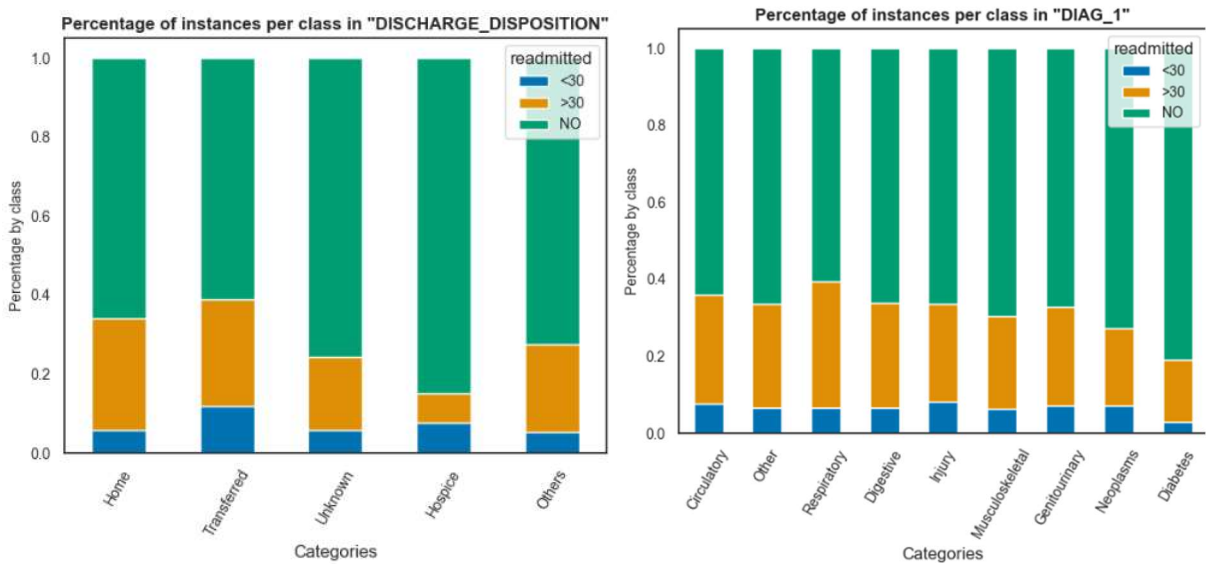

Figura 22. Porcentaje de muestras por clase y categoría en el "Conjunto Diabetes".

Adicionalmente, realizamos un análisis conjunto de las variables numéricas y categóricas mediante FAMD. En la *Figura 23* podemos observar aquellas variables que, a priori, parecen presentar redundancia, como es el caos de los pares 6 y 4 ('admission_type' y 'admission_source') y 18 y 19 ('change' y 'diabetesMed').

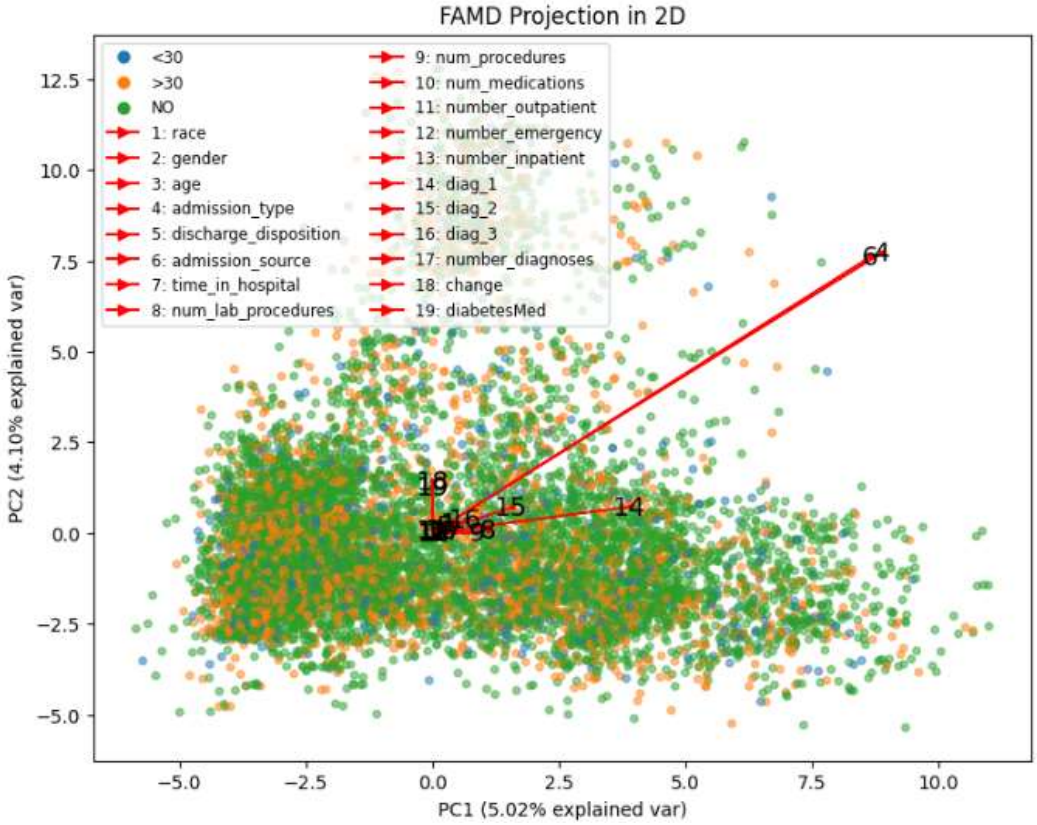

Figura 23. FAMD de 10000 puntos para todas las características del "Conjunto Diabetes"

A continuación, previo a la fase de entrenamiento, describimos los pasos de diseño, es decir los modelos a emplear, la separación de datos de entrenamiento y validación, las posibles codificaciones de datos, así como cualquier información relevante durante el diseño de los modelos.

Tras conocer el diseño de los modelos, podemos comenzar con el entrenamiento para posteriormente realizar su validación. En este caso, para observar los rendimientos empleamos **gráficos explicables durante el desarrollo**. Primero, generamos una curva ROC multiclase por modelo (*Figura 24*). En esta figura se muestra como los modelos presentan mejor rendimiento para la clase '<30'.

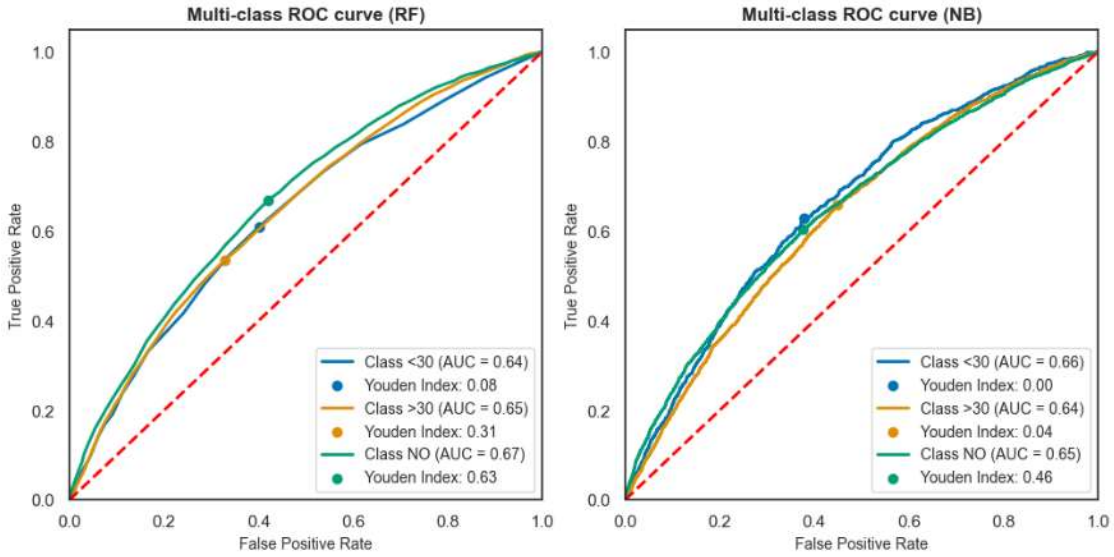

Figura 24. *Curvas ROC multiclase para modelos Random Forest y Naive Bayes en el "Conjunto Diabetes".*

Así pues, se puede observar como todos los modelos tienen un área bajo la curva muy similar para todas las clases. Sin embargo, estos gráficos no otorgan información muy concreta de las predicciones realizadas, por lo que, en la *Figura 25*, presentamos las matrices de confusión de cada modelo:

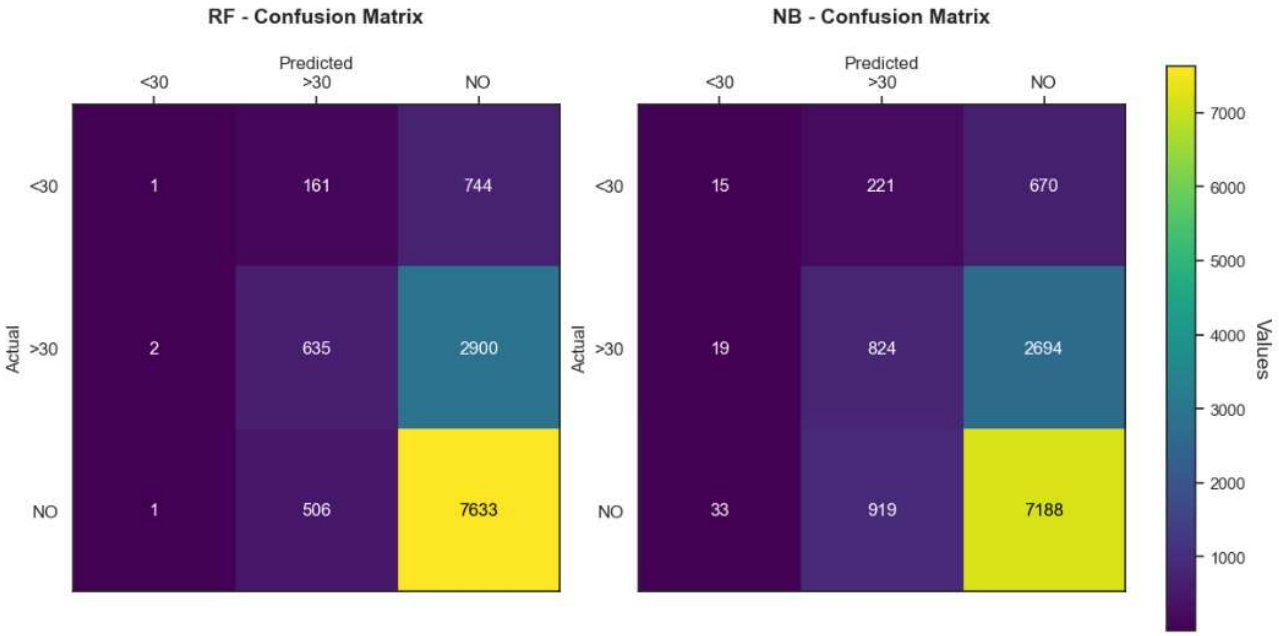

Figura 25. *Matrices de confusión para los modelos Random Forest y Naive Bayes en el "Conjunto Diabetes".*

De ellas extraemos con claridad cuáles son las predicciones reales, deduciendo que es considerablemente poco efectivo para la clase '<30', problema que intentaremos solventar en la siguiente etapa de Robustez técnica y seguridad.

Para poder entender cómo se están tomando las decisiones podemos evaluar en las predicciones, con posterioridad al desarrollo del modelo, cuál es la importancia de cada variable. Empleamos así pues **gráficos explicables post entrenamiento**. Esto se puede realizar para el problema multiclase, obteniéndose en la *Figura 26* un diagrama de barras apiladas según la importancia.

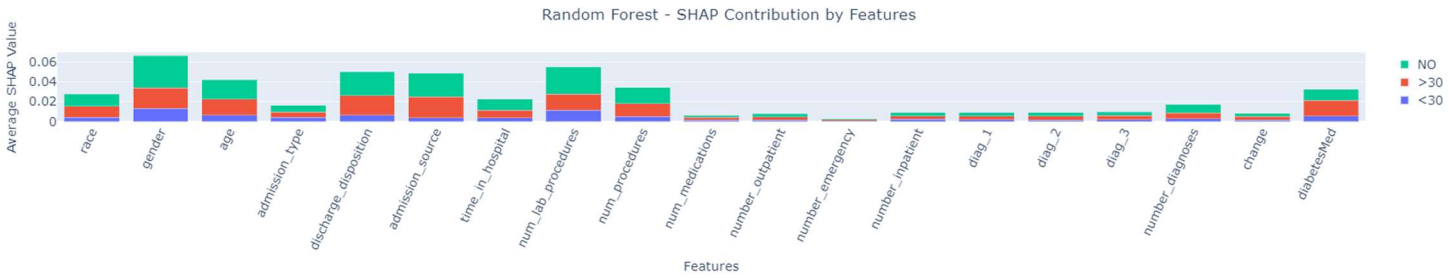

Figura 26. *Diagrama de barras apiladas con importancia de las variables según clase para el modelo Random Forest en el "Conjunto Diabetes".*

Y también se puede hacer para un problema binario estableciendo como '1' la clase positiva y '0' el resto. Con ello, obtenemos de nuevo un diagrama de barras apiladas, pero aparece la posibilidad de aplicar otros gráficos más informativos y a su vez complejos, como es el caso del gráfico derecho de la *Figura 27* correspondiente a un *beeswarm* de un modelo de Regresión Logística.

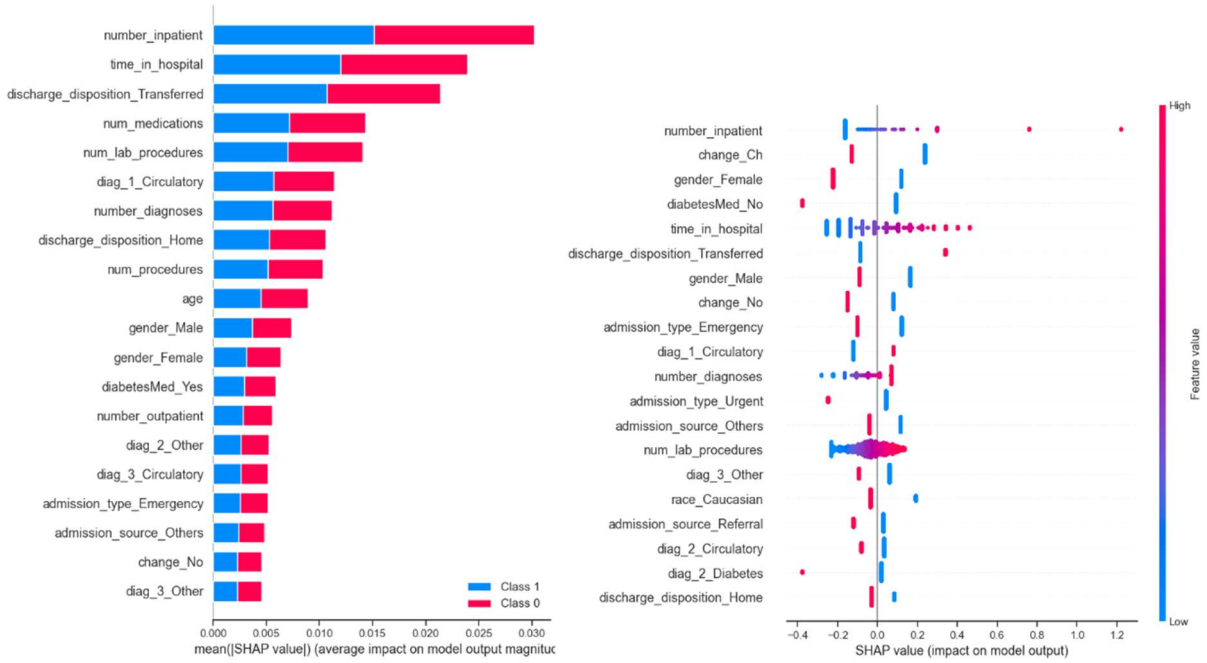

Figura 27. *Gráficos explicativos de la predicción para una clasificación binaria en el "Conjunto Diabetes".*

Tras los gráficos explicativos, podemos cumplimentar la plantilla de *disclaimer* proporcionada del siguiente modo:

DISCLAIMER FOR TRUSTWORTHY AI MODEL RANDOM FOREST

This trustworthy AI model, ***Random Forest V1****, has been meticulously developed to* ***predict diabetic readmission****s. It underwent extensive training on* ***around 100 000 instances spanning10 years*** *and consistently demonstrates a high level of accuracy, with performance metrics indicating* ***0.64 AUC ROC*** *when assessed in* ***the same context using test data****. The model has undergone rigorous testing for fairness and robustness, although it's important to acknowledge* ***that there may be bias between races and genders[optional].***

Random Forest V1 *employs inherently* ***not transparent*** *algorithms to the end-user, and we* ***employ SHAP graphics*** *to ensure transparency and interpretability. Data privacy and security are paramount, and we uphold strict measures****, such as change logs****, to safeguard user data in compliance with regulations.*

While ***Random Forest V1*** *has been carefully designed to minimize biases and ethical concerns, safeguarding patient's safety, all in line with the EU AI Act and EU Medical Device Regulation users are urged not to rely solely on it for* ***critical decisions****. It is imperative to exercise caution when interpreting the model's predictions and, when appropriate, consult with a qualified human expert for review.*

To maintain your trust in ***Random Forest V1****, we are dedicated to routine updates and maintenance, ensuring it remains aligned with the latest data and best practices. Should you have any inquiries, concerns, or encounter any issues, please do not hesitate to reach out to us at bdslab@upv.es.*

Please take note that ***BDSLab*** *cannot be held liable for any harm or damage resulting from the use of* ***Random Forest V1*** *outside of its designated use cases and recommendations.*

Además, en la *Tabla 2* ejemplificamos una de documentación de salida de un entrenamiento guardando métricas como la precisión, especificidad o puntuación F1 en un archivo que contenga la fecha y hora de ejecución:

Tabla 2. *Ejemplo de registro de las métricas de un entrenamiento del modelo Random Forest para el "Conjunto Diabetes"*

Precision	Recall	F1 Score	Support	Model version	Date
0.59	0.66	0.58	12583	Version 1.0	2024-06-19 09:14:08

4.3.5. Robustez técnica y seguridad

Con los datos ya acondicionados y una vez contamos con una metodología establecida para comprender el modelo, podemos comenzar a mejorar su rendimiento y seguridad. Cabe destacar que la ejemplificación de este requisito se realizará binarizando el problema, ya que las clasificaciones multiclase comprenden una casuística algo más infrecuente que limita en cierta medida el uso de algunos métodos.

Para la aplicación en cuestión no vamos a realizar una selección de características formal. De existir una gran redundancia entre variables se habría observado en la etapa de Transparencia por medio de grandes valores de correlación. En su lugar, recomendamos, tras obtener la importancia individual de cada variable, eliminar las características de menor importancia analizando la constancia del rendimiento para identificar el punto de convergencia del modelo. Así pues, entramos directamente al diseño y desarrollo del modelo. Con el objetivo de realizar un entrenamiento insesgado, el **balanceo de clases** mediante *SMOTE* y la **optimización de hiperparámetros** con *GridSearch* se hacen de forma conjunta dentro de cada bloque del método de CV (con el número de bloques k=5), imprescindible para realizar una **evaluación bien formada**. Una vez contamos con el umbral de saturación óptimo (Índice de Youden), podemos realizar una validación con todos los datos de entrenamiento para observar el rendimiento genérico. Observamos en la *Figura 28* las curvas ROC de ambos modelos.

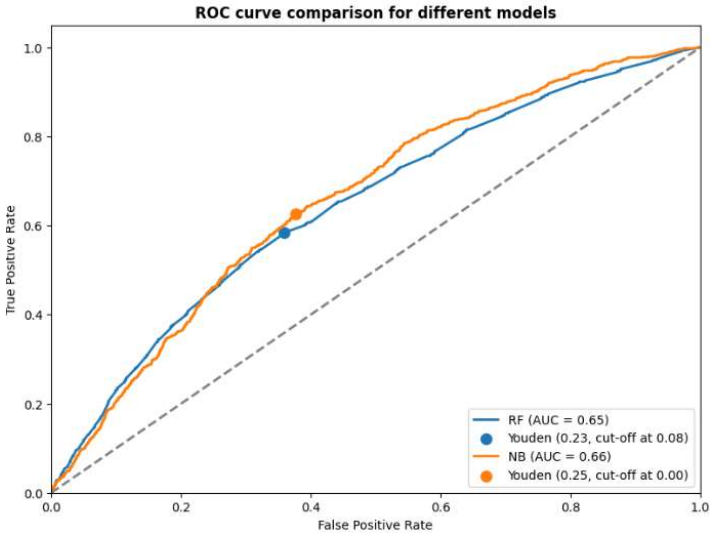

Figura 28. Curvas ROC para diferentes modelos entrenados de forma óptima y balanceada en el "Conjunto Diabetes".

Para comparar con los resultados obtenidos sin optimizar el rendimiento, en la *Figura 29* observamos las matrices de confusión resultantes.

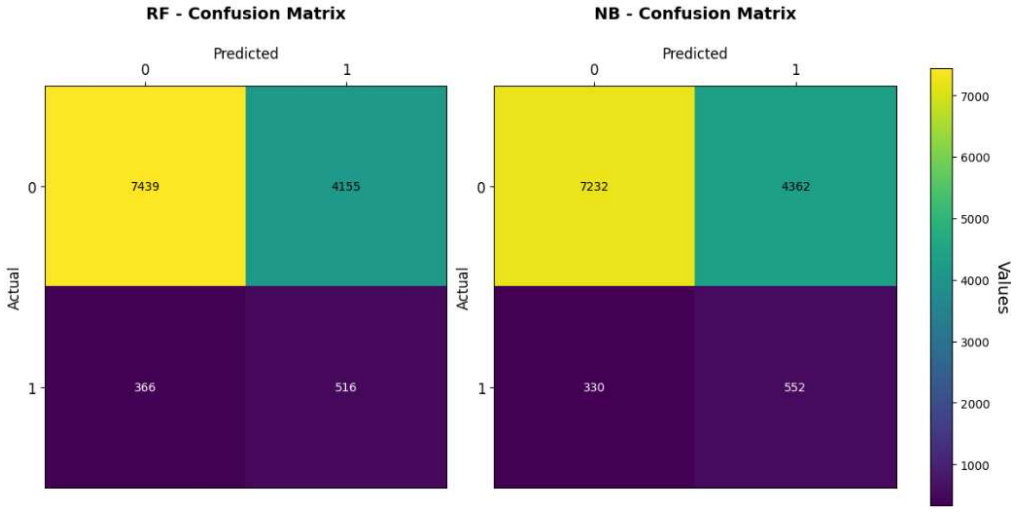

Figura 29. Matrices de confusión para varios modelos entrenados de forma robusta en el "Conjunto Diabetes"

Del mismo modo, podemos utilizar otras herramientas como las curvas de calibración para complementar la validación de los diferentes modelos y obtener otra perspectiva del comportamiento. En la *Figura 30* apreciamos el resultado de obtener la curva de calibración. Idealmente buscaríamos que la línea de puntos azul se ajustará a la diagonal discontinua de color gris, para lo cual también deberíamos de tener las distribuciones de probabilidad separadas.

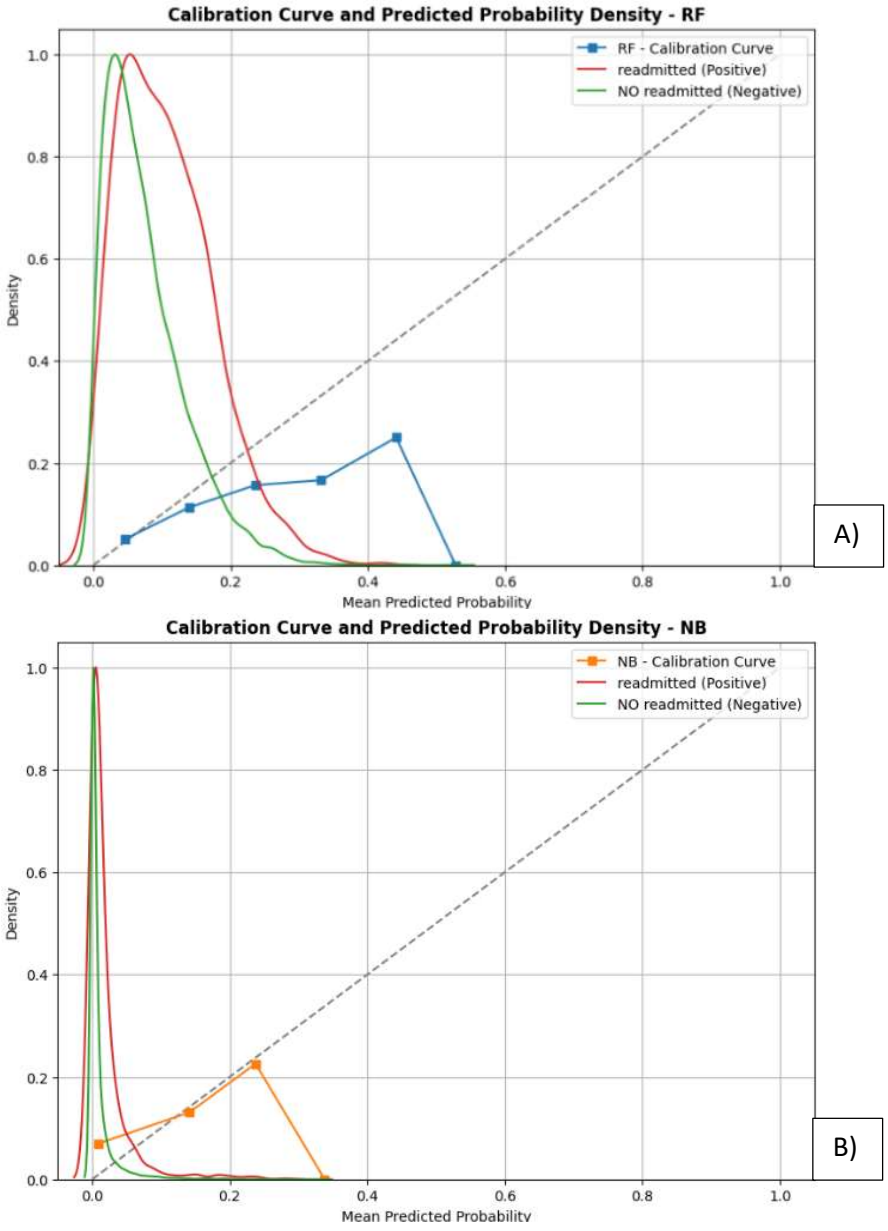

Figura 30. *Curvas de calibración y densidad de probabilidad para diferentes modelos en el "Conjunto Diabetes": A) Modelo Random Forest, B) Modelo Naive Bayes.*

Para cuantificar la incertidumbre del modelo, nos valemos de un método *Bootstrap* para entrenar 100 modelos con distintos parámetros y un 30 % de los datos de entrenamiento con reemplazo. Esto nos permite conocer la media y desviación típica de las métricas del modelo (Véase la *Tabla 3*).

Tabla 3. Métricas conjuntas de 100 modelos Bootstrap para el "Conjunto Diabetes".

	Random Forest			Naive Bayes		
	Sensibilidad	**Especificidad**	**F1-score**	**Sensibilidad**	**Especificidad**	**F1-score**
Promedio	42.59 %	70.52 %	75.92 %	49.49 %	67.51 %	74.36 %
Desviación típica	8.38 %	6.55 %	4.05 %	5.55 %	4.7 %	2.98 %

Posteriormente si aplicamos los modelos Bootstrap entrenados en la etapa anterior para **la cuantificación de la incertidumbre en las predicciones**, podemos obtener con el conjunto de los modelos la respuesta de todos ellos para la predicción de dos muestras, un caso con clase 0 y otro con clase 1 (*Figura 31*). Para ambos modelos, el caso 1 presenta una distribución de probabilidad con mucha variabilidad pero con valores mayormente situados en la sección superior al umbral de saturación de cada modelo.

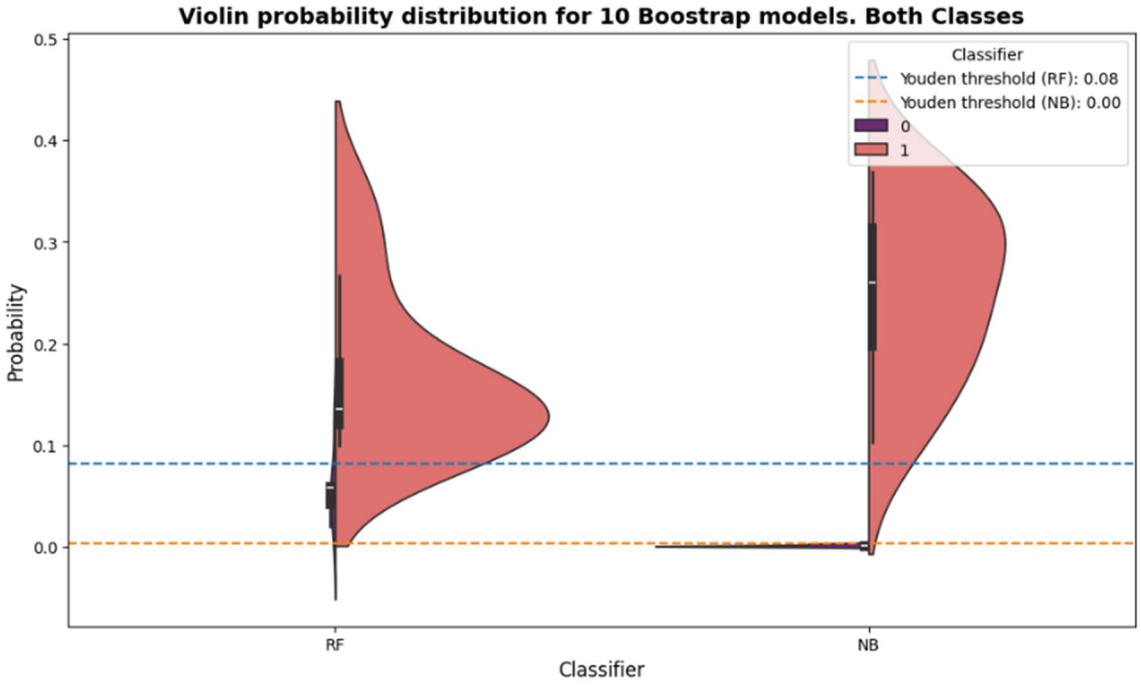

Figura 31. Distribución de probabilidad de los modelos Bootstrap para distintos datos del "Conjunto Diabetes".

De un modo similar, si perturbamos una muestra e intentamos predecir los nuevos datos obtenidos, es posible estimar el rendimiento frente a pequeñas variabilidades. En este punto podemos analizar tanto la variabilidad general de la predicción como de la clase, identificando mediante el sesgo si la distribución total es ancha o estrecha, así como si contiene al Índice de Youden o por el contrario se encuentra alejada. En la *Figura 32* mostramos dos ejemplos, uno donde se ofrecería un resultado de precisión confiable (A), ya que el umbral de saturación no se encuentra dentro del intervalo de confianza de la distribución de probabilidad, y otro no confiable (B) al encontrar el umbral dentro del intervalo:

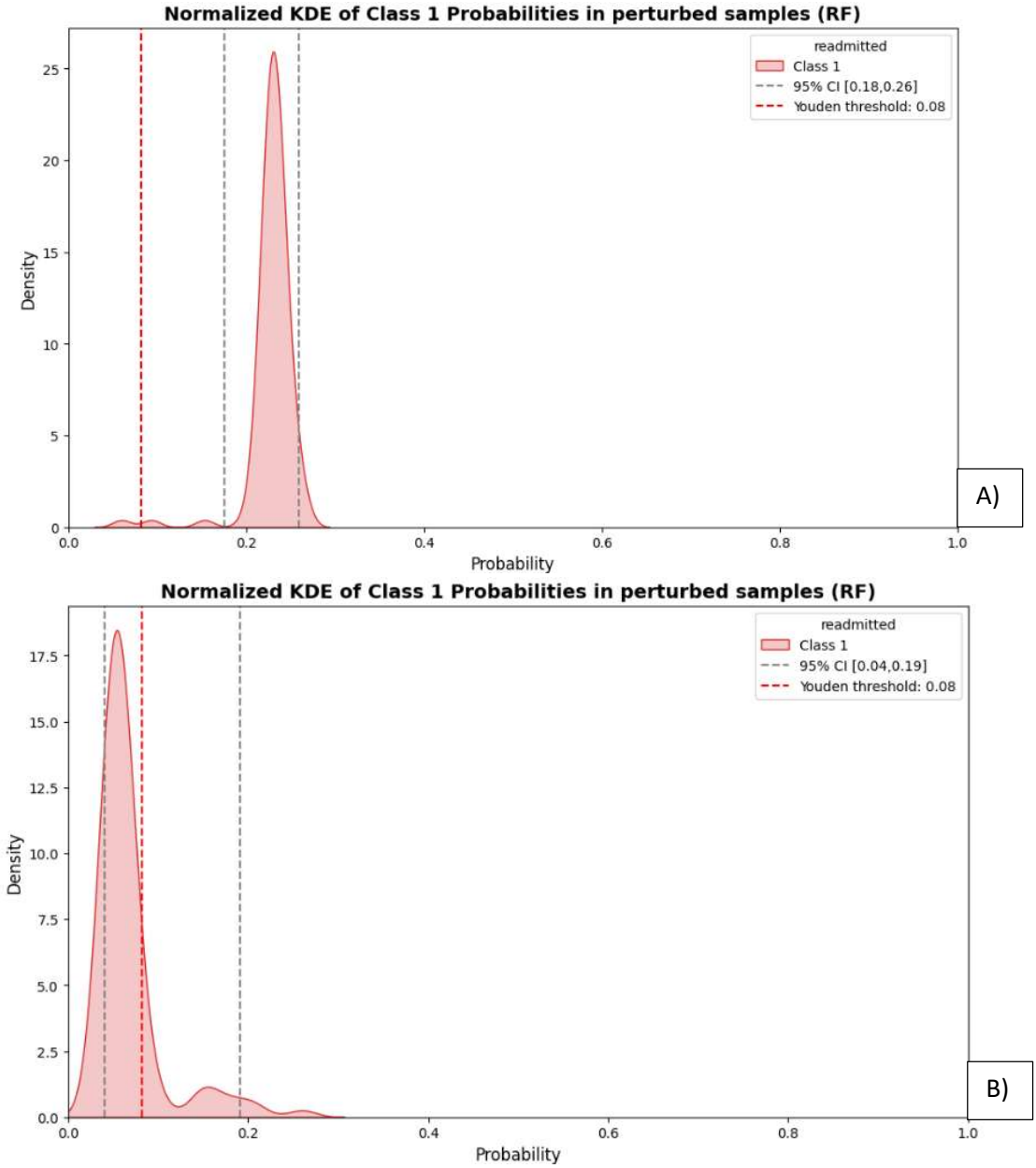

Figura 32. *Densidades de probabilidad de dos casos positivos perturbados del "Conjunto Diabetes": A) Predicción segura, B) Predicción insegura.*

Seguidamente, probamos también cuál es la respuesta del modelo frente a los datos que hemos detectado como outliers en la etapa de Privacidad y Gobernanza de datos, ya que puede ser un indicador interesante del manejo que tiene el sistema predictivo frente a los datos especialmente complejos de clasificar. En la *Figura 33* realizamos una comparación entre los métodos Bayesiano (*Bootstrap*) y de Aumentado para datos anómalos y normales. De ella concluimos que, en general, y especialmente para los modelos Bootstrap, la distribución de probabilidad de los outliers presenta una mayor varianza.

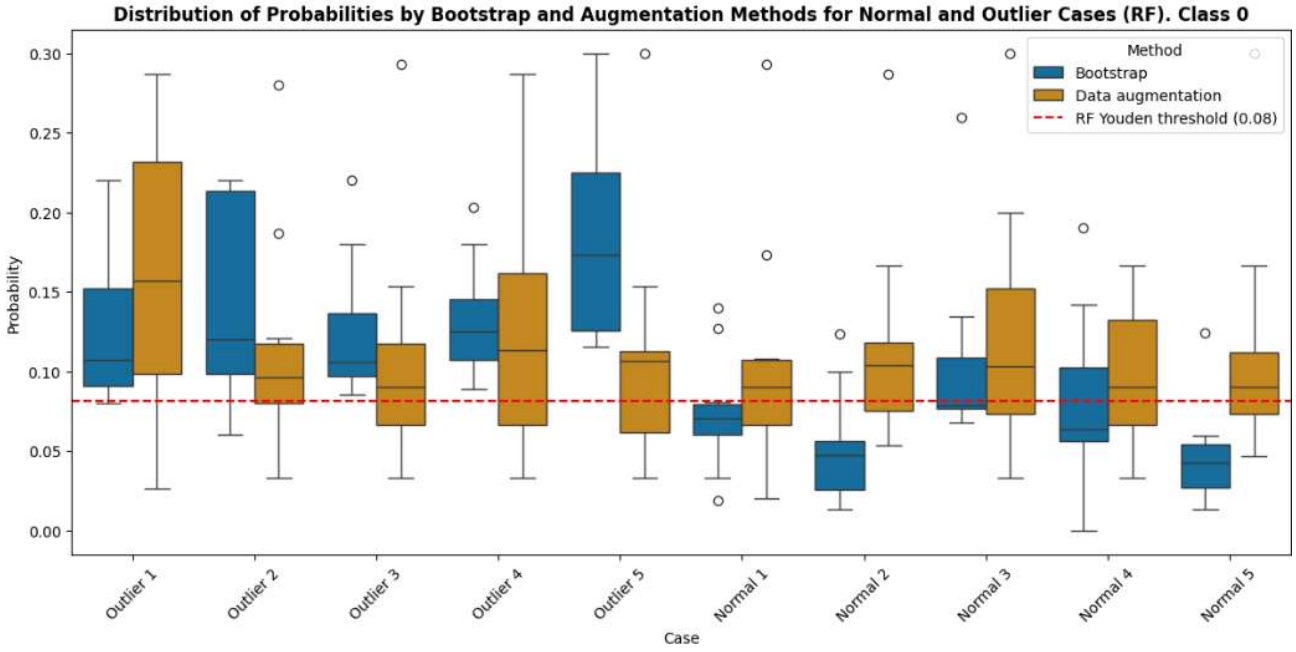

Figura 33. Distribuciones de probabilidad para datos anómalos y normales predichos con varios modelos bootstrap o perturbados del "Conjunto Diabetes".

Por último, empleando las fechas determinadas anteriormente de forma simulada, ejemplificamos cómo podría realizarse una monitorización de la variación temporal del rendimiento de un modelo, con el fin de detectar ***dataset shifts***. En la *Tabla 4* mostramos las métricas resultantes de entrenar un modelo 1 con los datos de 1999-2004 y predecir 2005-2006, y entrenar un modelo 2 con 1999-2006 y predecir 2007-2008.

Tabla 4. Métricas para modelos entrenados con variabilidad temporal

	Random Forest	
	Sensibilidad	**F1-score (promediado macro)**
Modelo 1 | Entrenado con datos de 1999-2004	37 %	52 %
Modelo 2 | Entrenado con datos de 1999-2006	41 %	54 %

4.4. Evaluación en conjunto de datos “Heart Disease”

A continuación, mostramos los resultados de evaluar los pipelines propuestos para el “conjunto Heart Disease”.

4.4.1. Metadatos

El “conjunto Heart Disease” es considerablemente más sencillo, cuenta con un número menos de muestras, características y variedad en general. Así pues, las únicas transformaciones previas que realizamos son las conversiones de codificación ordinal de las variables categóricas, que por defecto vienen como números, a tipo objeto, ya que en el interior del pipeline realizaremos la codificación pertinente.

Así pues, en la siguiente *Tabla 5* podemos apreciar los metadatos correspondientes al “conjunto Heart Disease”. Destacamos en ellos la falta de clase positiva, ya que el problema de clasificación es en sí binario. Aun así, emplearemos por consistencia una clase positiva de valor ‘1’. Además, el conjunto de datos carece de variables identificativas.

Tabla 5. Metadatos iniciales para el “conjunto Heart Disease"

METADATOS PARA EL “CONJUNTO HEART DISEASE”	
Conjunto de datos	'dataset_heart_disease_full.xlsx'
Salida	‘target’
Clase positiva	‘ 1‘
Características identificativas	‘ ‘
Características sensibles	‘sex’
Característica para balancear	‘sex’
Proveniencia de los datos	[“El conjunto de datos consta de 1190 registros de pacientes de EE. UU., Reino Unido, Suiza y Hungría.”, ‘ ‘]
Fecha de adquisición	[“Sin información”, ‘ ‘]
Tipos de características	▪ age: numérica, ▪ sex: categórica, ▪ chest pain type: categórica , ▪ resting bp s: numérica , ▪ cholesterol: numérica , ▪ fasting blood sugar: categórica , ▪ resting ecg: categórica , ▪ max heart rate: numérica , ▪ exercise angina: categórica ", ▪ oldpeak: numérica , ▪ ST slope: categórica ,

4.4.2. Privacidad y gobernanza de datos

En este caso, la calidad de los datos (**control de la calidad de los datos**) es considerablemente mayor. A pesar de que se ajustan en menor medida a la realidad por el volumen y la actualidad, presentan una mayor corrección y completitud. En la *Figura 34* podemos observar cómo tan solo constan datos perdido en la variable “colesterol”.

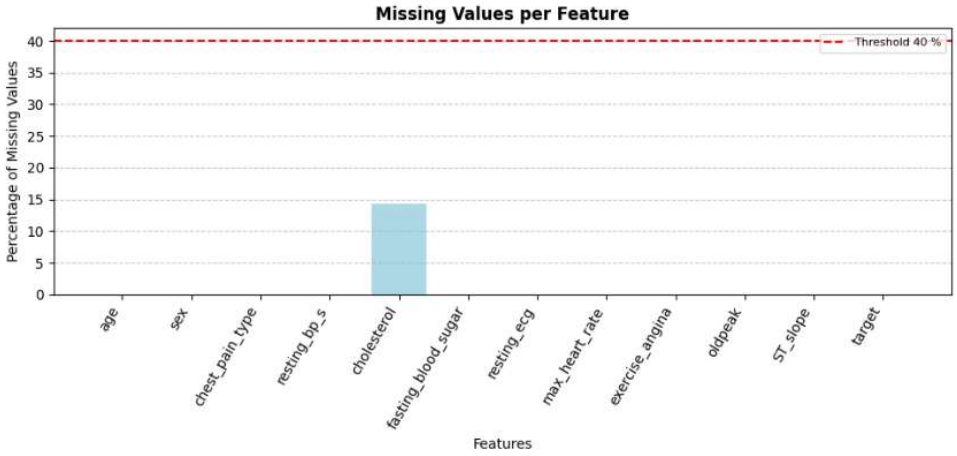

Figura 34. *Gráfico de barras de valores perdidos con umbral de eliminación en 40% perdidos en el "Conjunto Heart Disease".*

En este caso, no tenemos tampoco *outliers* categóricos, ya que todas las categorías tienen al menos un 1% del volumen de datos total. En cuanto a los *outliers* univariantes numéricos, aplicando un percentil 98 se detectan 95 celdas anómalas, lo cual corresponde a un 1'78% del total de celdas numéricas. Aplicamos los mismos métodos de imputación que en "Conjunto Diabetes", una regresión lineal y KNN con 10 imputaciones.

Por último, realizando el mismo estudio conjunto de PCA y HTA, para un nivel de significancia α=0.01 no se detecta ningún *outlier* multivariante. En la *Figura 35* observamos la proyección en 3D de los datos con las nuevas componentes principales.

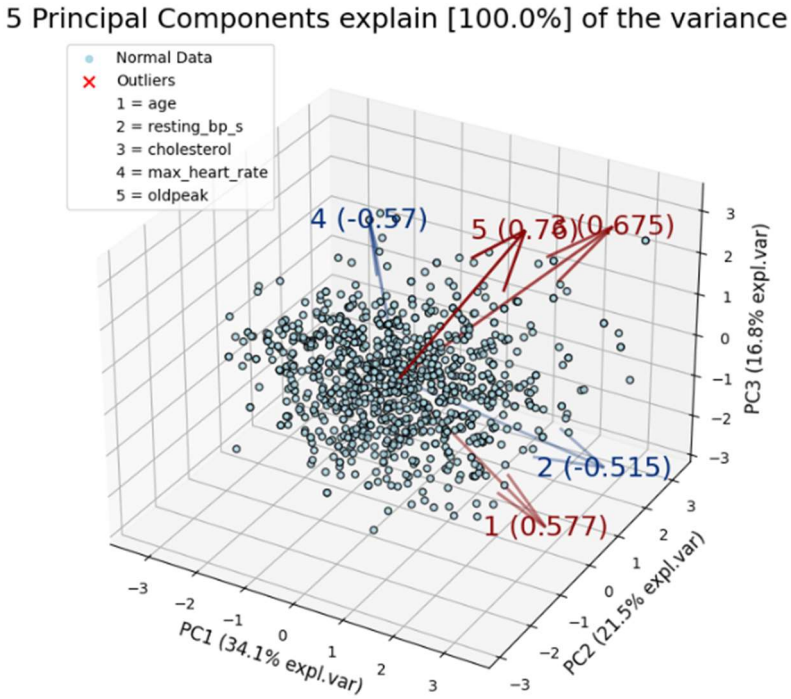

Figura 35*. Proyección de las 3 componentes principales tras realizar un PCA de las características numéricas del "Conjunto Heart Disease".*

4.4.3. Diversidad, no discriminación y justicia

Del mismo modo, realizamos un **análisis exploratorio sensible**. En primer lugar, en la *Figura 36* podemos visualizar la frecuencia absoluta según los diferentes sexos y clases.

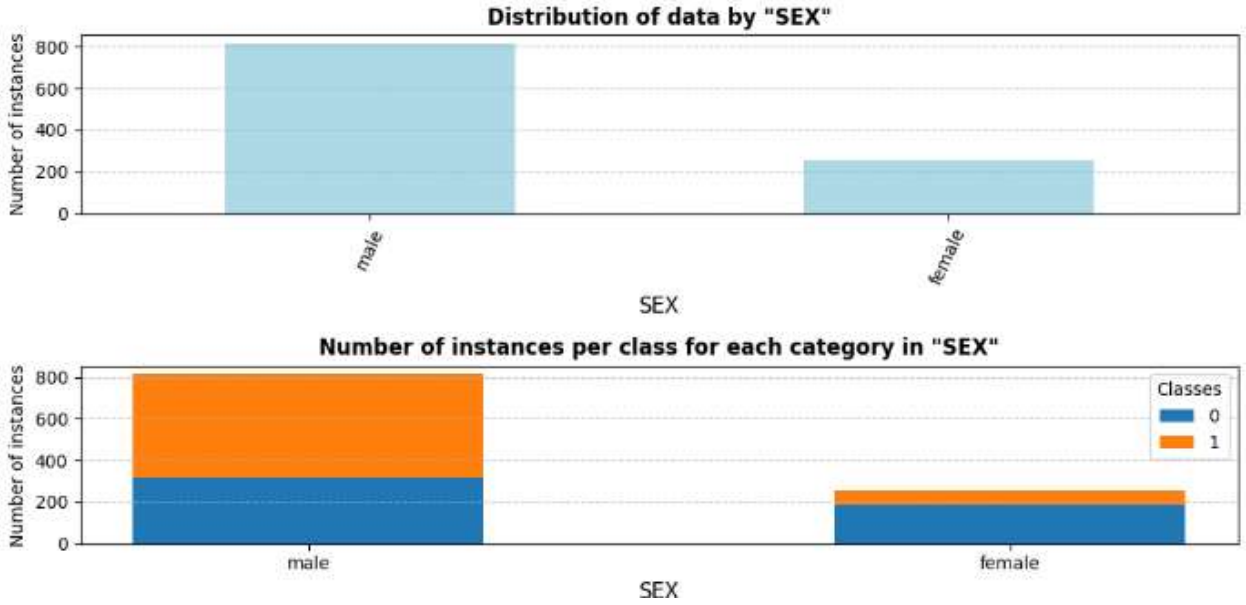

Figura 36. Distribución de datos por categoría y clase para cada variable sensible en el "Conjunto Heart Disease".

En cuanto a la distribución de datos perdidos, en la *Figura 37* se aprecia una mayor densidad de perdidos en los hombres, lo cual es relativamente lógico en vista de la cantidad de datos de cada sexo.

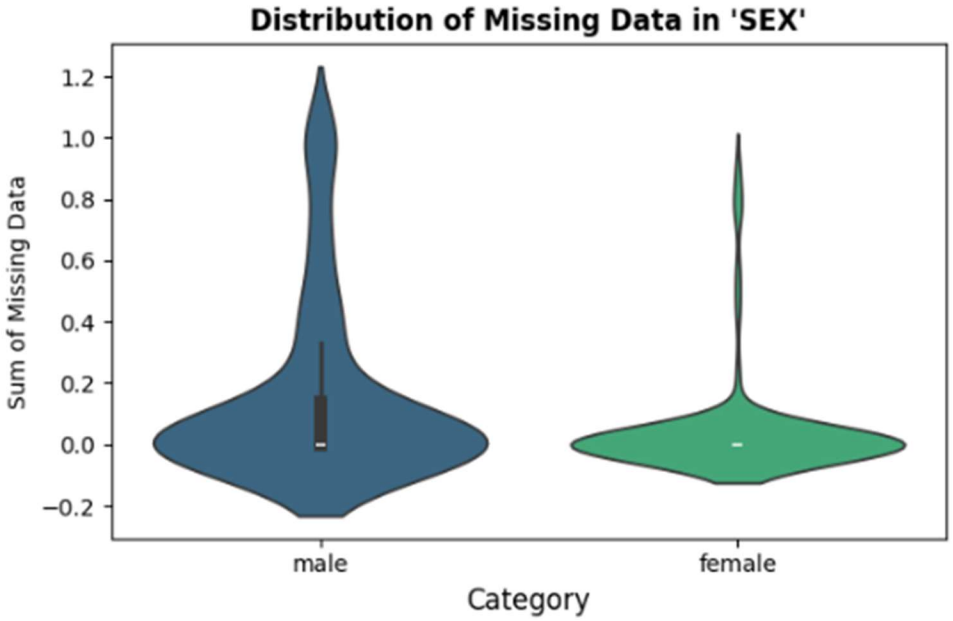

Figura 37. Distribución de los datos perdidos según el sexo

En cuanto a las diferencias de correlación entre categorías sensibles, tan solo contamos con la diferencia entre sexos, apreciable mediante el mapa de calor de la *Figura 38*.

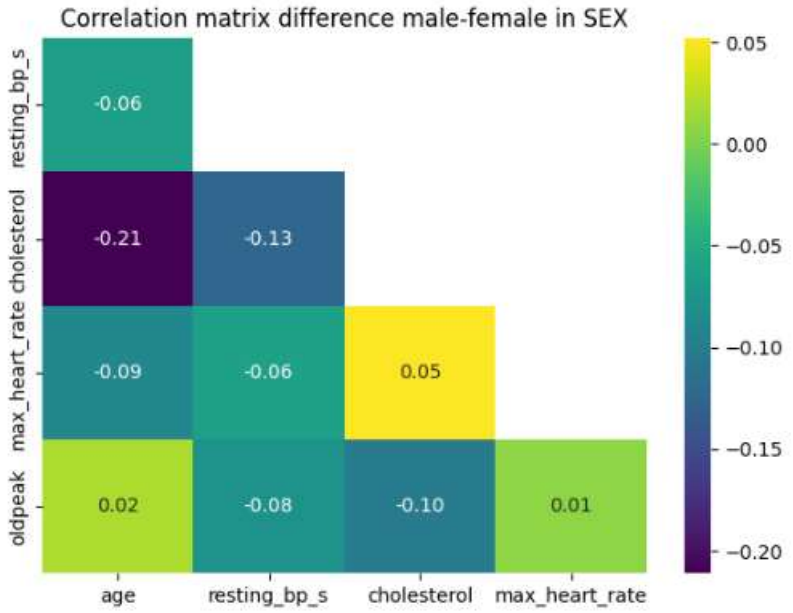

Figura 38. Mapa de calor de la diferencia de las matrices de confusión obtenidas con las categorías 'male' y 'female' en la variable sensible 'sex' *en el "Conjunto Heart Disease".*

Si evaluamos ejecutamos las posibles rutas de acción frente a la existencia de sesgos, obtenemos la *Figura 39*, donde tenemos la comparación de los tres métodos.

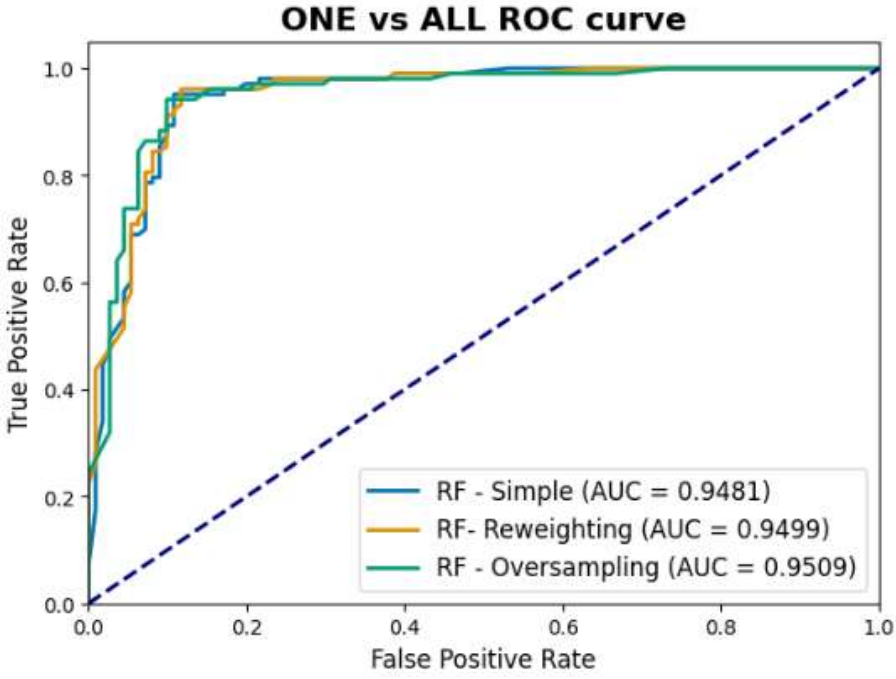

Figura 39. *Comparación curvas ROC para un modelo simple, con reponderación y con sobremuestreo en el "Conjunto Heart Disease".*

En el caso de predecir datos de un sexo particular, obtenemos curvas ROC por sexo (**Rendimiento por variables sensibles**). En la *Figura 40* apreciamos el resultado donde parece que, si no se mitiga el sesgo, el modelo muestra peores predicciones para las mujeres.

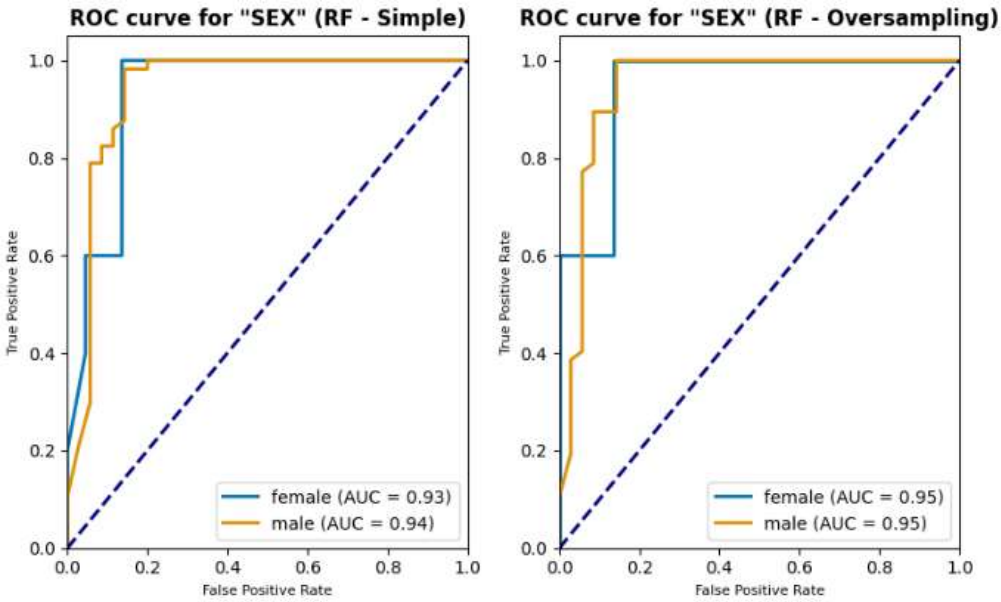

Figura 40. Comparación curvas ROC de modelos entrenados con muestras de diferente sexo en el "Conjunto Heart Disease".

También podemos entrenar modelos únicamente con datos de un sexo particular para observar la importancia de las variables. En la *Figura 41* observamos el resultado de los modelos entrenados para cada categoría, donde para las variables 'exercise_angina' y 'ST_slope' se aprecian distancias considerables.

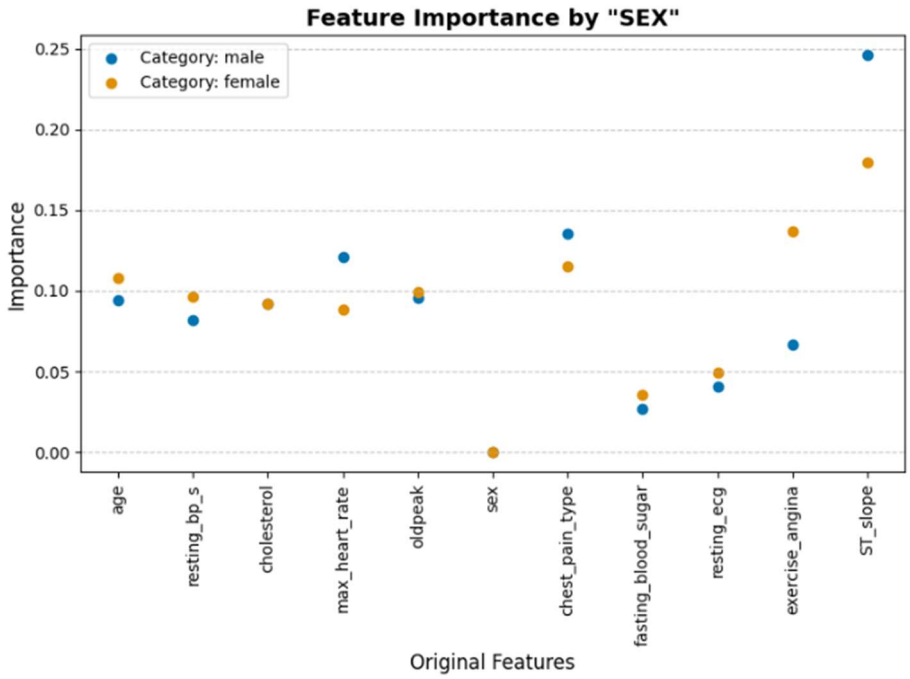

Figura 41. Importancia de las variables para modelos Random Forest entrenados con distintas categorías

4.4.4. Transparencia

En este caso, al contar con menos datos, puede realizarse el **análisis exploratorio general** de los datos con un gráfico bivariante para los datos numéricos (Véase *Figura 42*).

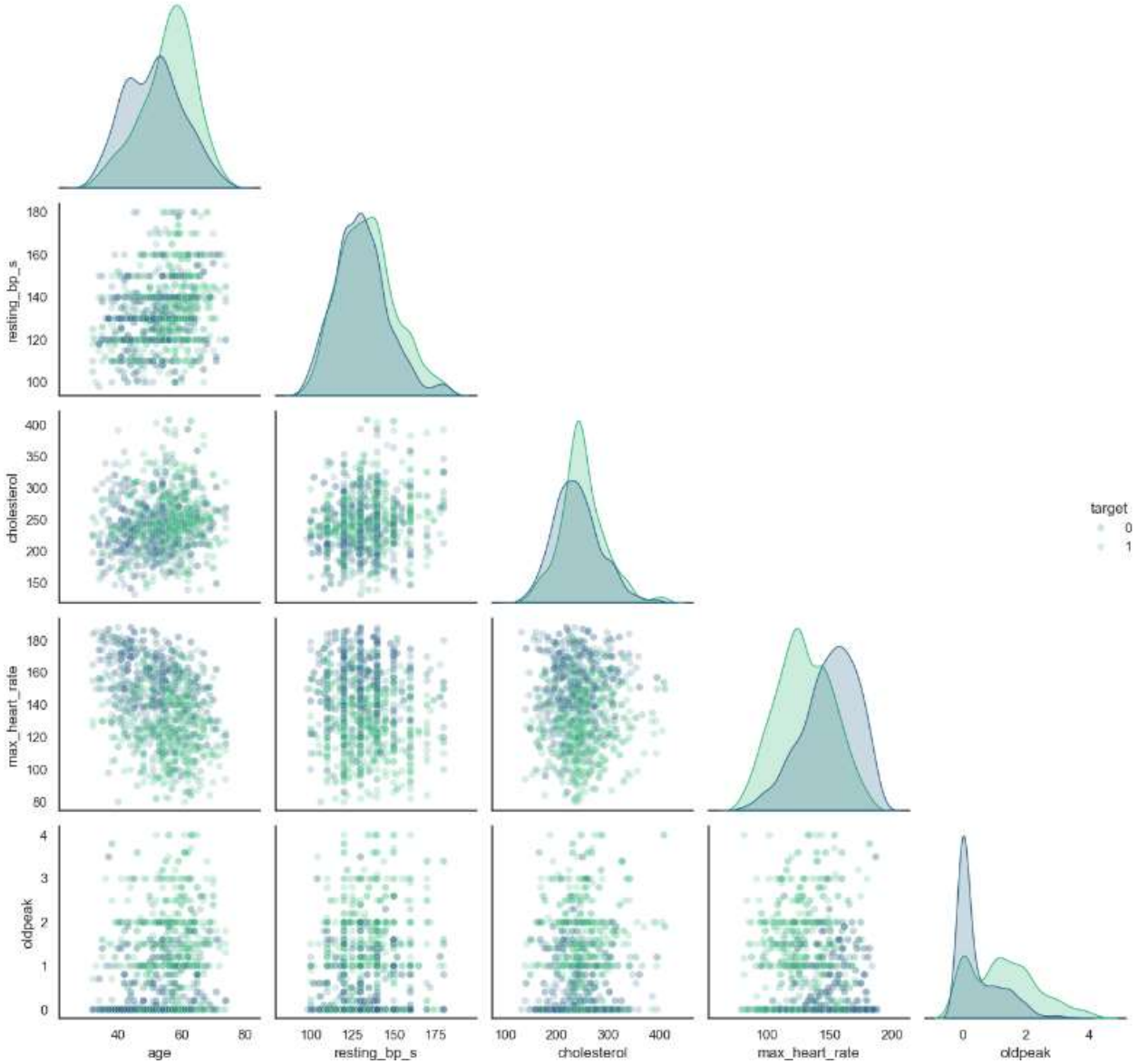

Figura 42. *Gráfico bivariante de las características numéricas en el "Conjunto Heart Disease".*

Al igual que con el "Conjunto Diabetes", graficamos la proporción de clase por categoría del "Conjunto Heart Disease" en la *Figura 43*, donde parece que las variables son más discriminativas ya que las presentan diferente frecuencia relativa según las categorías.

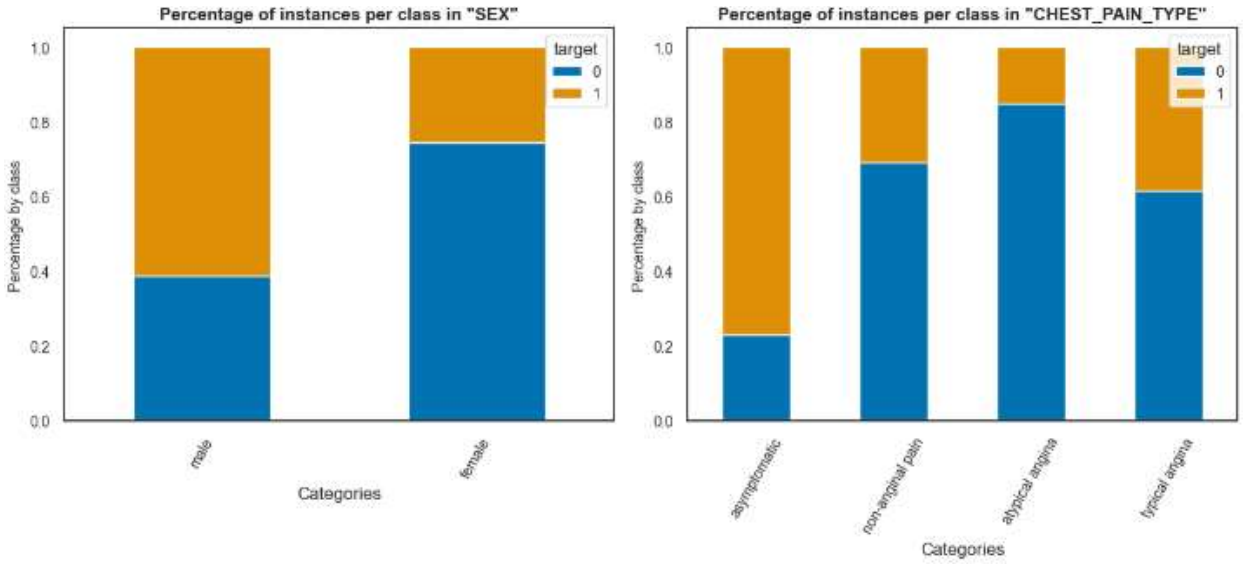

Figura 43. Porcentaje de muestras por clase y categoría en el "Conjunto Heart Disease".

Seguidamente, el análisis categórico y numérico conjunto con FAMD muestra un análisis aparentemente más claro, donde las nuevas componentes principales separan considerablemente las clases (*Figura 44*).

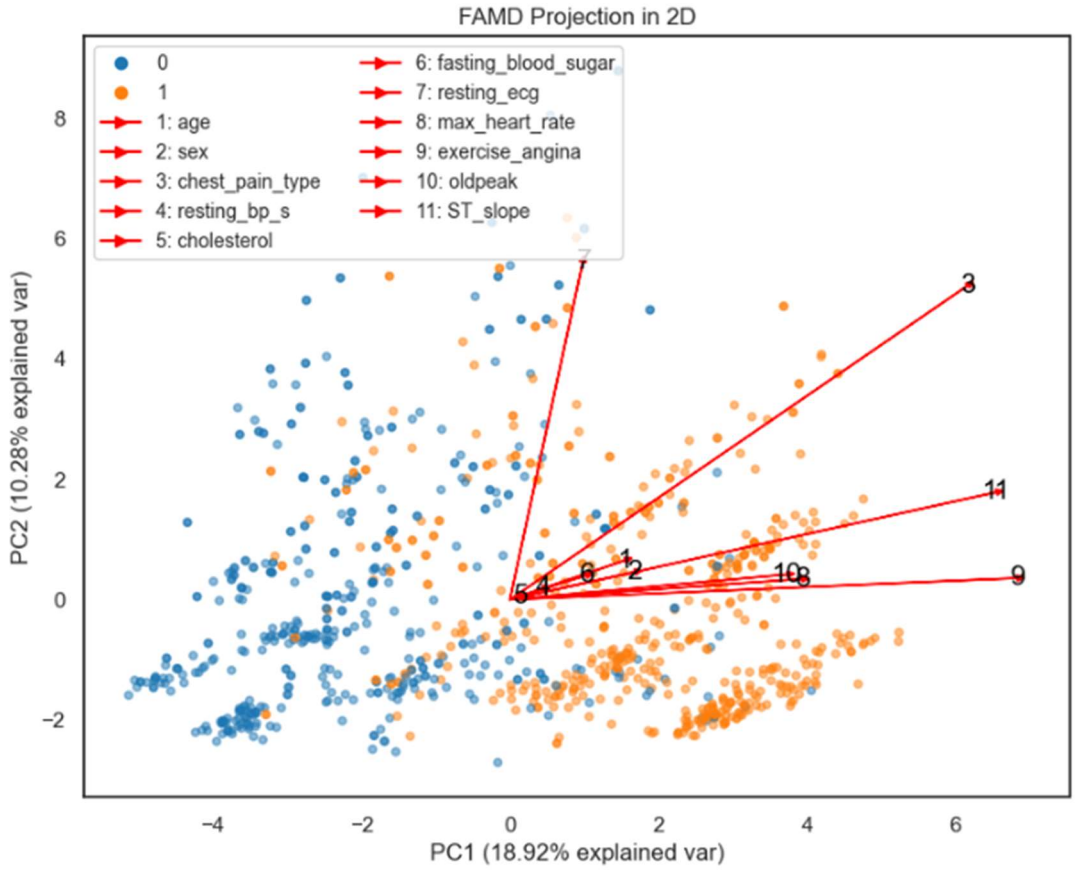

Figura 44. Proyección 2D de FAMD para el "Conjunto Heart Disease"

A continuación, tras el entrenamiento, caben esperar mejores resultados que para el conjunto anterior. En la *Figura 45* observamos la matriz de confusión resultante (**Gráficos explicables**).

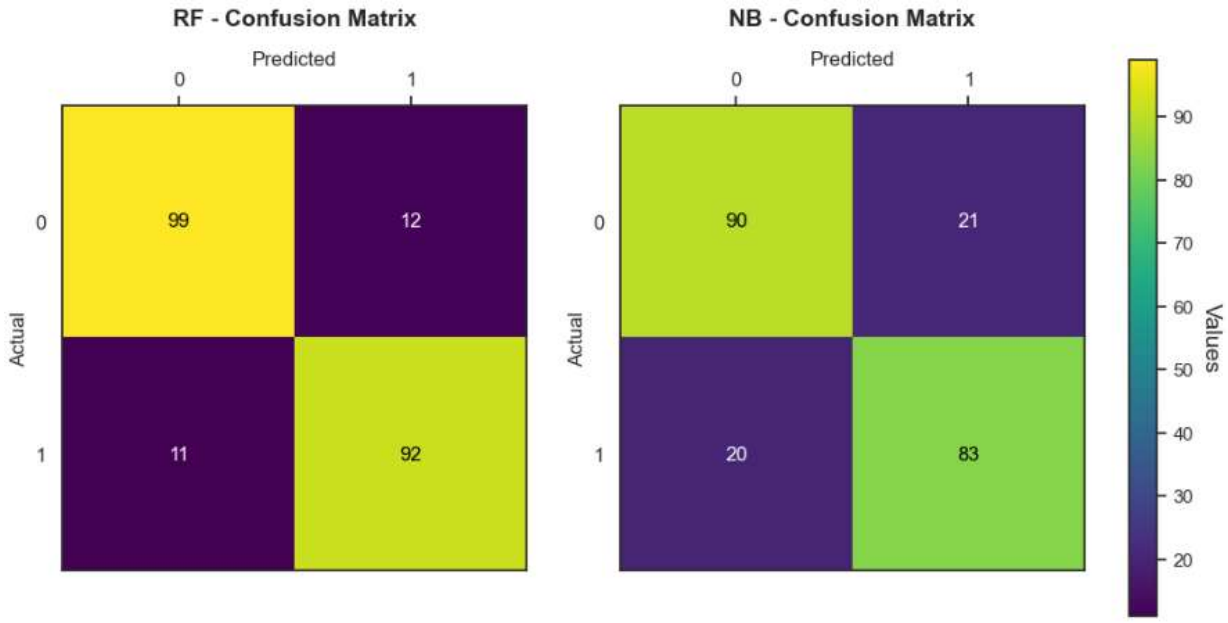

Figura 45. Matrices de confusión para los modelos Random Forest y Naive Bayes *en el "Conjunto Heart Disease".*

4.4.5. Robustez técnica y seguridad

Tras realizar la **optimización de hiperparámetros**, el **balanceo de clases** y el entrenamiento pertinente mediante validación cruzada **(Evaluación bien formada**), obtenemos las curvas ROC de la *Figura 46*.

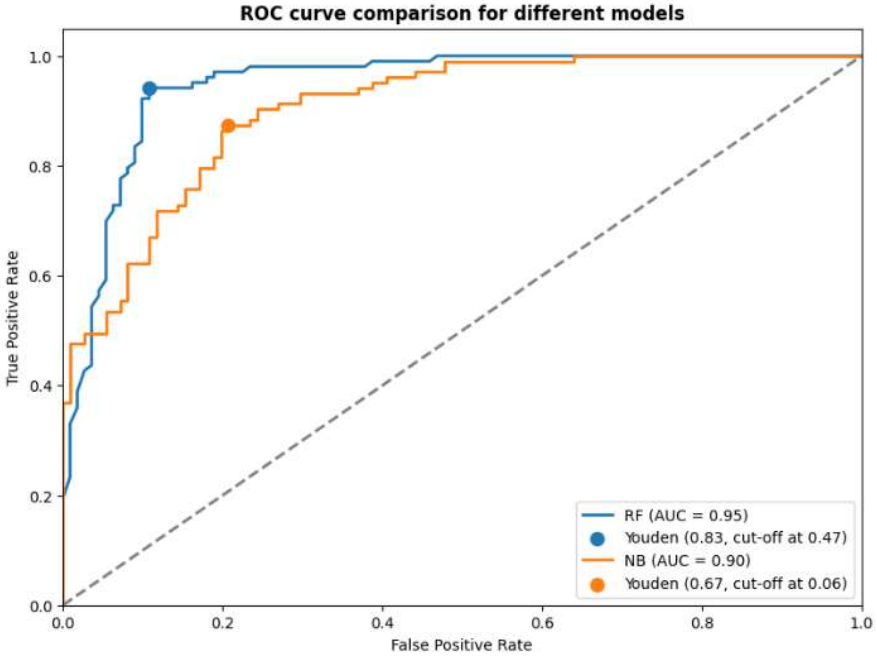

Figura 46. *Comparación de las curvas ROC para Random Forest y Naive Bayes tras optimizar hiperparámetros y balancear las clases en el "Conjunto Heart Disease".*

Como cabe esperar, las curvas de calibración son también mejores (*Figura 47*).

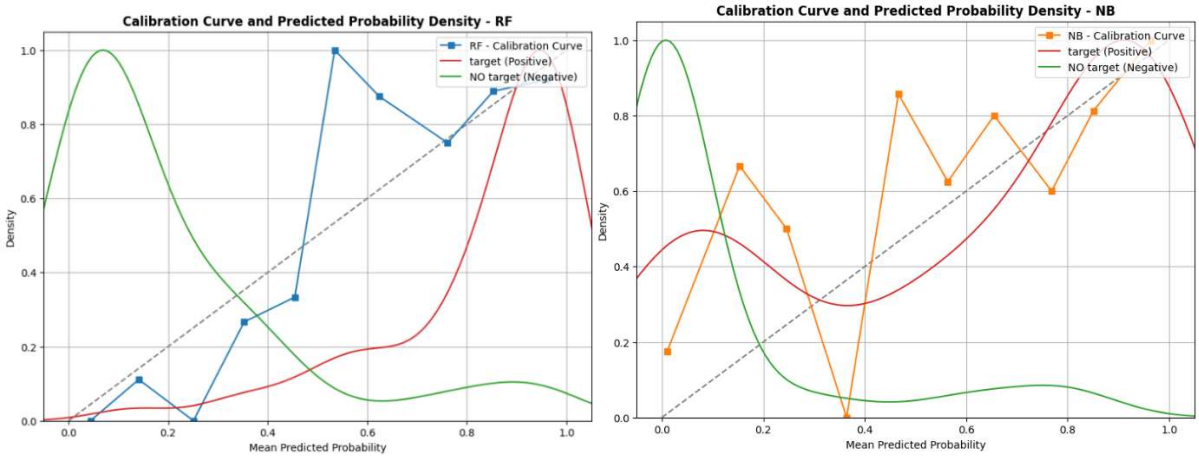

Figura 47. *Curvas de calibración y densidad de probabilidad para los diferentes modelos en el "Conjunto Heart Disease".*

Los modelos Bootstrap, de un modo similar al conjunto anterior, presentan distribuciones de probabilidad para la clase positiva superiores a los respectivos umbrales de para cada modelo (*Figura 48*).

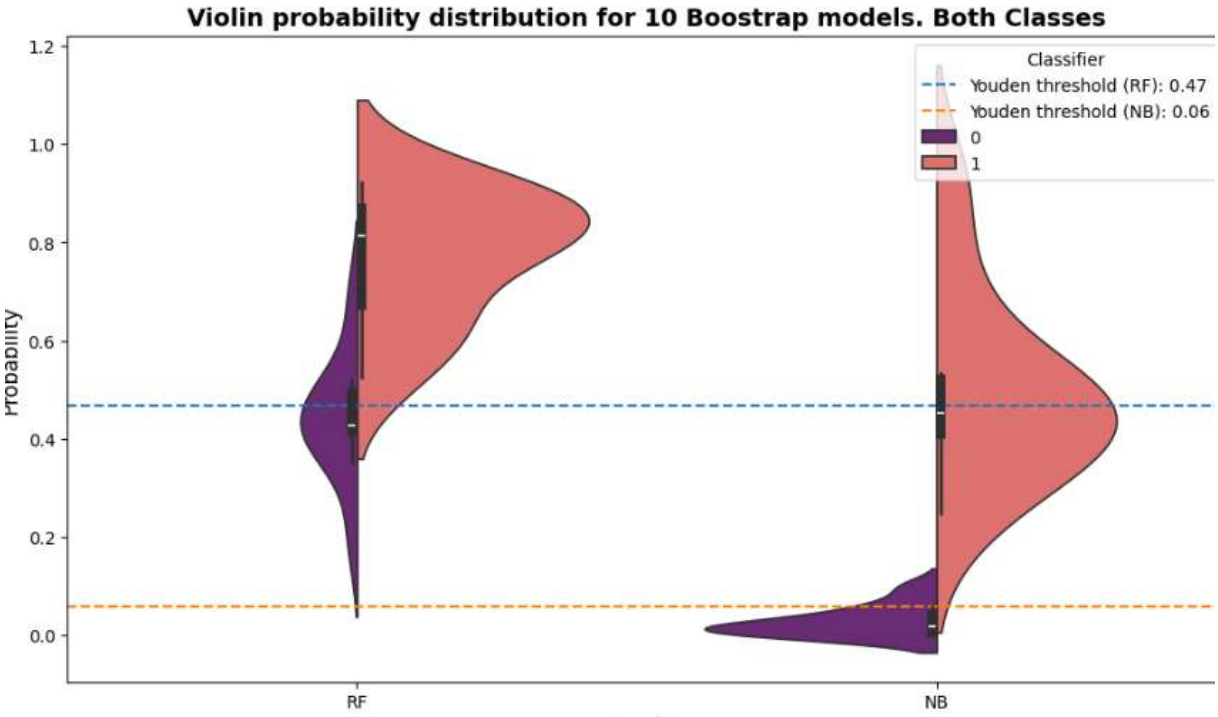

Figura 48. *Distribución de probabilidad de los modelos Bootstrap para distintos datos del "Conjunto Heart Disease".*

Por último, al perturbar las muestras, detectamos que la mayoría de los casos presentan una distribución de probabilidad claramente distanciada del umbral de saturación (*Figura 49*), por lo que las predicciones, en este contexto, serían seguras.

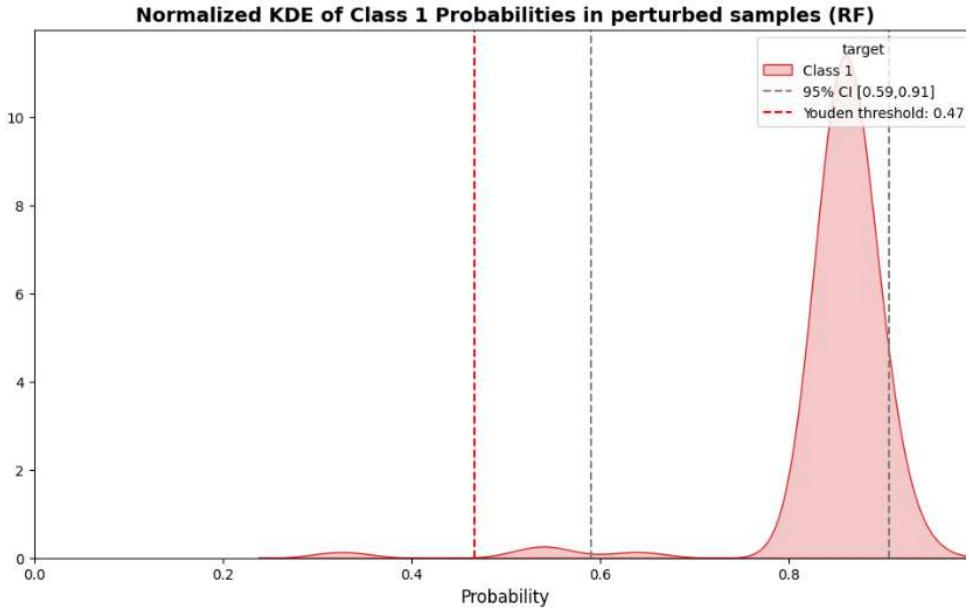

Figura 49. *Densidad de probabilidad de un caso positivo perturbado del "Conjunto Heart Disease".*

4.5. Checklist de recomendaciones para el desarrollo de IA confiable

A partir de los resultados anteriores podemos elaborar una *checklist* genérica que concentra todos los aspectos a considerar para alcanzar una IA confiable que satisfaga todos los requisitos establecidos por la UE. El resultado en cuestión se aprecia en la *Tabla 6*, donde se incluyen todos los requisitos abordados, así como los componentes y métodos técnicos que los apoyan. La *checklist* queda abierta a ampliaciones a medida que se añadan métodos técnicos nuevos para satisfacer los requisitos no cubiertos en esta investigación o para mejorar los existentes, como es el caso por ejemplo de métodos de prevención de ataques.

Además, a modo de soporte, en la versión online incluimos en cada componente un hipervínculo para acceder de manera rápida a los resultados del "conjunto Diabetes" donde se podrá encontrar la sección específica en la que se trata.

Tabla 6. Checklist de recomendaciones para una IA confiable según requisito y ciclo de vida (P: Data preparation, D: Model development, U: Deployment & Use, M: Management, N/A: Not applicable)

CHECKLIST OF TRUSTWORTHY AI RECOMMENDATIONS ***based on EU guidelines*** (European Comission, 2019)				
Stage	**Requirement**	**Item**	**Checklist item**	**Satisfied**
	Data collection and metadata			
N/A	Data types and ranges	M1a	Data types have been identified	
		M1b	The valid ranges or domain for each variable are defined	
	Data provenance	M2	Metadata includes information about the origin of data	
	Target variable	M3a	The target variable is defined	
		M3b	The positive class is defined	
	Sensitive Variables	M4a	The possible sensitive variables have been identified and provided	
		M4b	A sensitive variable has been selected to balance its instances	
	Identification Variables	M4	The identification variables are defined	
	Privacy and data governance			
P	Anonymization	G1	The data does not contain any identifying information	
	Standardisation	G2	The information is adapted to the standards of the context	
	Data quality control	G3a	A data quality assessment has been carried out based on well-defined dimensions	
		G3b	Data variability across data sources and over time has been evaluated	
M	Changelogs	G4	A log system changes exists	
	Diversity, non-discrimination and fairness			
P	Sensitive exploratory analysis	F1	A sensitive exploratory analysis has been done to find possible relationships associated with sensitive variables in the dataset	
	Bias mitigation	F2a	There is implemented a bias mitigation pre-processing method	
D		F2b	There is implemented a bias mitigation in-processing method	
U		F2c	There is implemented a bias mitigation post-processing method	
D	Sensitive variables performance	F3	Model performance is evaluated specifically for each sensitive variable	
U	Fairness monitoring	F4	The system includes a fairness monitoring system in the deployment	
	Transparency			
P	General exploratory analysis	T1	A general exploratory analysis has been done to find possible relationships or further problems in the dataset variables	
D	Design description	T2	System design steps are considerably explained	
	Explainability plots	T3a	Explainability techniques or visualizations are included to support the results of the training model.	
U		T3b	Explainability techniques or visualizations are included to support the results of the deployed model	
	Disclaimer	T4	The system includes a disclaimer with its features and limitations	
M	Output Documentation	T5	Outputs are carefully recorded	
	Incident Sharing	T6	There has been provided a way to report incidents	
	Technical robustness and safety			
P	Dimensionality	R1	Potential dimensionality problems are handled with feature selection or dimensionality reduction methods	
	Class balancing	R2a	There is implemented a class balancing pre-processing method	
		R2b	There is implemented a class balancing in-processing method	
D	Hyperparameter optimization	R3	Extensive hyperparameter optimization has been performed to optimize model performance	
	Well-formed evaluation	R4	Metrics evaluation provides robust results appropriate to the context	
	Attack prevention	R5	There have been implemented methods to prevent the model for attacks of different nature	
	Uncertainty Quantification	R6a	Uncertainty is quantified and reported in metric	
U		R6b	Uncertainty is quantified in predictions	
		R6c	Predictions have a threshold of uncertainty above which are not made	
	Dataset shift monitoring	R7	Dataset shifts are considered and handled in the development & monitored in further use	
	Retraining	R8	A retraining system for external implementations has been considered.	

** Item structure: Requirement – Index – Subsection Index | EXAMPLE: T3a (Requirement: Transparency – Index: 3 – Subsection: a)*

*** This checklist is open to new methods for the current and remaining requirements*

CAPÍTULO 5. Discusión

5.1. Impacto

En primer lugar, cabe destacar que este trabajo permitiría el desarrollo de una IA que tendría la confianza y seguridad requerida en los entornos de salud, donde decisiones basadas en datos podrían tener impacto directo en millones europeos. Además, al estar alineado con recomendaciones europeas como son la guía de la IA confiable (European Comission, 2019) y Ley de la IA (European Parliament, 2024) facilitaría la certificación de los mismos sistemas, especialmente cuando son de alto riesgo, ámbito en el cual se incluyen los sistemas basados en salud.

Seguidamente, el impacto de la investigación es relevante tanto en su capacidad de generalización como en su aplicabilidad en conjuntos de datos. Se ha mostrado como es un esquema idóneo a partir del cual un Ingeniero de Datos o cualquier otro profesional especializado en el área puede comenzar a construir su sistema de IA asegurando su confiabilidad.

Notamos que es una propuesta genérica, donde para realizar la experimentación con nuevos conjuntos de datos prácticamente no precisa de cambios, pero a su vez muestra sensibilidad a la dificultad de los datos, mostrando cómo funciona a la perfección conjuntos de datos ideales, como es el caso del "conjunto Heart Disease", y mostrando resultados adaptados a la realidad de los datos para el "conjunto Diabetes". Cabe destacar que la metodología se presenta como un pipeline semiautomático. Sin embargo, la propuesta comprende aplicaciones genéricas, por lo que para obtener el mayor rendimiento se debería, a partir de las directrices aportadas, adaptar los métodos y funcionalidades al contexto particular de implementación del sistema de IA confiable.

5.2. Posicionamiento en el estado del arte

El presente trabajo se posiciona a la vanguardia del estado del arte en inteligencia artificial confiable. A diferencia de las metodologías tradicionales que se enfocan exclusivamente en aspectos específicos de la confiabilidad como los marcos de trabajo expuestos por Deloitte (Deloitte, 2022) o IBM (IBM, 2021), esta propuesta integra de manera holística un conjunto de métodos que abordan los requisitos establecidos por la Unión Europea para garantizar la confianza en un sistema de IA. La tecnicidad de los métodos es también en si misma un avance respecto de la guía de la UE (European Comission, 2019), la cual supone la referencia principal.

Así pues, a diferencia de otros enfoques segmentados del estado del arte cuyo foco se reduce al aspectos concretos, como TRIPOD-AI (Collins et al., 2024) que se centra en el desarrollo o en la validación del modelo, la presente metodología resulta extremadamente completa y rigurosa, ya que considera cada requisito a lo largo de todo el ciclo de vida, examinando todos los aspectos derivados del desarrollo de la IA desde el comienzo de la recolección de los datos hasta la implementación y gestión del modelo. Asimismo, su adaptabilidad y aplicabilidad suponen una solución integral y cubren un vacío significativo en el campo de la IA confiable, correspondiendo a una de las pocas metodologías que incorporan métodos técnicos y ejemplificaciones en el abordaje de la construcción del sistema

5.3. Limitaciones y trabajo futuro

Para comenzar, es importante destacar que el alcance del proyecto es muy elevado, por lo que nos hemos focalizado en unos aspectos principales, dejando algunas técnicas o secciones para trabajo futuro, como pueden ser las señaladas en la *checklist*.

Continuando la línea de alcance, a pesar de los avances logrados con el desarrollo del pipeline, podemos percibir la existencia de ciertas carencias inherentes al diseño y los aspectos abarcados. Es este el caso de la falta de tratamiento de datos en formato de texto libre o imágenes. En un contexto con este tipo de información, a pesar de que los principios fundamentales serían equivalentes, los métodos técnicos aplicados podrían diferir ligeramente. Por alcance temporal no se ha tratado con este tipo de datos, sin embargo, como hemos dicho, la metodología queda abierta por lo que es un aspecto para considerar en trabajo futuro.

En esta misma línea, el *pipeline* se ha diseñado de modo semiautomático debido a la complejidad de la automatización, principalmente causada por el alta combinatoria de acciones a realizar. En consecuencia, la metodología requiere una intervención humana que, pese a ser escasa en comparación con un diseño totalmente de cero, puede introducir variabilidad y potenciales sesgos y, en cualquier caso, supone una falta de agilidad y una dificultad añadida. En cualquier caso, los *pipelines* se han diseñado de forma suficientemente genérica para que, a partir de la definición de metadatos inicial, la intervención humana en el resto de código para los distintos principios sea lo menor posible.

Así pues, estas limitaciones apuntan hacia unas áreas de mejora y expansión del sistema basadas en el manejo y la aceptación de una mayor variedad de datos, así como la realización de una automatización completa de la construcción del sistema, pudiendo además añadir nuevas técnicas de IA confiable. Por tanto, si continuamos con el trabajo futuro que emana de las limitaciones encontradas, la dirección clave se enfocará en la mejora y expansión de las capacidades del pipeline.

En primer lugar, una prioridad en el desarrollo es la automatización. Resulta fundamental desarrollar una interfaz intuitiva y completamente automatizada, mediante la cual cualquier usuario pueda cargar un conjunto de datos y, sin trabajo adicional, recibir de manera directa todos los resultados necesarios para evaluar la confiabilidad del sistema. De manera similar, se podría evolucionar la matriz de inteligencia artificial confiable para convertirla en dinámica e interactiva, permitiendo modificar los métodos existentes.

Por otro lado, sería de vital importancia mejorar el rendimiento y la generalización del pipeline, para lo cual sería especialmente interesante emplear redes neuronales. El estudio se ha hecho con modelos de aprendizaje automático sencillos; sin embargo, aplicar técnicas de aprendizaje profundo podría mejorar los resultados y la generalización de la metodología. De hecho, esta podría ser una vía para abarcar toda la gama de datos existentes, siendo las redes neuronales una herramienta prometedora para generalizar hacia datos de tipo texto libre e imágenes.

En un segundo plano, encontraríamos la necesidad de aplicar métodos técnicos que consideren los requisitos aun no abordados formalmente: Agencia humana y supervisión, Bienestar social y medioambiental, y Responsabilidad. Como se explicó en capítulos anteriores, la Unión Europea no establece ningún tipo de jerarquía entre requisitos , por lo que debe considerase igual de relevante el cumplimiento de estos tres respecto a los cubiertos en este trabajo. Sin embargo, si bien es cierto que

no hemos definido métodos específicos para estos requisitos, hemos detectado que presentan cierta transversalidad y pueden tener áreas parcialmente cubiertas. Es el caso por ejemplo, de la Agencia humana y supervisión, y Bienestar social y medioambiental, ya que ambas se encuentran ampliamente vinculadas con la Diversidad, no discriminación y justicia y los métodos de mitigación de sesgo estarían favoreciendo el cumplimiento de los dos requisitos. Del mismo modo, el requisito de Responsabilidad cuyo objetivo es asegurarse que se cumplen el resto de requisitos y buscar responsables en vista de problemas, podemos afirmar que, a pesar de no contar con registros formales para ello, el mero estricto cumplimiento del resto de requisitos supondría la consideración de una primera área de este requisito.

Otro aspecto referente a la ampliación de la guía es el aumento y mejora de los metadatos con el fin de elevar la calidad del estudio. En relación con el Espacio Europeo de Datos Sanitarios (EHDS, por sus siglas en inglés), que establece un marco para el intercambio seguro y eficiente de datos de salud en toda la UE, el enriquecimiento de los metadatos no solo permitiría una mejor estandarización e interoperabilidad de los datos entre diferentes sistemas y países, sino que también proporcionaría un contexto más detallado sobre las condiciones de recolección y las características técnicas de los datos. Además, al añadir más información contextual, se facilitaría la reproducción y validación de los resultados, incrementando la transparencia y la trazabilidad del estudio.

Finalmente, sería interesante analizar la utilidad real de la metodología con una cantidad de usuarios significativa. Si bien es cierto que ya se ha validado el código por tres expertos en ciencia de datos, la investigación no deja de ser una propuesta teórica complementada con aspectos prácticos que necesita de un mayor reporte de los usuarios finales. Cabe mencionar también que la guía se encuentra online por lo que es susceptible de recomendaciones por parte de futuros terceros usuarios para seguir mejorando. Con el código en abierto se podrá considerar el criterio de un mayor número de personas susceptibles de utilizar el pipeline para detectar los potenciales puntos de fricción existentes.

CAPITULO 6. Conclusiones

El presente proyecto ha satisfecho de manera efectiva todos los objetivos propuestos. Se ha logrado establecer un *pipeline* que sirve como acercamiento a la creación de una inteligencia artificial confiable en el ámbito de la salud. Cumpliendo con el objetivo inicial, se ha conseguido desarrollar una guía metodológica orientativa de las acciones a realizar para desarrollar, implementar, utilizar y gestionar sistemas de IA confiables según los principios y gran parte de los requisitos establecidos por la Unión Europea (Privacidad y gobernanza de datos, Diversidad, no discriminación y justicia, Transparencia, y Robustez técnica y seguridad).

En relación con el objetivo específico 1 (O1) que buscaba especificar de manera técnica de principios y requisitos de la IA confiable a lo largo de su ciclo de vida, se han analizado las componentes y condiciones de manera exitosa, determinando posteriormente aquellos métodos técnicos capaces de satisfacer los requisitos. Además, se ha desarrollado una matriz que sirve como marco gráfico y mapa conceptual para relacionar los métodos con el ciclo de vida del sistema y los requisitos de confiabilidad, simplificando y facilitando la comprensión de los procesos.

El mayor logro de la investigación podría atribuirse al cumplimiento del objetivo específico 2 (O2), el cual comprende al desarrollo de las libretas dinámicas individuales para cada principio de confiabilidad, ofrecidas en código libre en https://github.com/bdslab-upv/trustworthy-ai. Las libretas actuales constituyen una herramienta práctica para implementar sistemas que se acerquen a la confiabilidad y seguridad.

Podemos confirmar, tras evaluar con varios conjuntos de datos del ámbito de la salud, que el pipeline puede llegar a ser generalizable, por lo que el objetivo específico (O3) también resultaría cubierto. Sin embargo, a pesar de que la intervención humana ha mostrado ser mínima, en la versión actual la supervisión humana es necesaria. Además, aun cuando el sistema abarca la mayoría de requisitos y ya ha sido evaluado por profesionales del sector, deberemos actualizar la metodología propuesta para los tres requisitos restantes.

En conclusión, este trabajo establece una base sólida para el desarrollo de sistemas de IA confiable en salud, proporcionando un esquema metodológico práctico y adaptable que, alineado con la ciencia abierta, puede ser expandido y mejorado para encontrar la solución óptima en cada contexto particular.

CAPÍTULO 7. Bibliografía

Abdar, M., Pourpanah, F., Hussain, S., Rezazadegan, D., Liu, L., Ghavamzadeh, M., Fieguth, P., Cao, X., Khosravi, A., Acharya, U. R., Makarenkov, V., & Nahavandi, S. (2021). A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information Fusion*, *76*, 243-297. https://doi.org/10.1016/j.inffus.2021.05.008

Albahri, A. S., Duhaim, A. M., Fadhel, M. A., Alnoor, A., Baqer, N. S., Alzubaidi, L., Albahri, O. S., Alamoodi, A. H., Bai, J., Salhi, A., Santamaría, J., Ouyang, C., Gupta, A., Gu, Y., & Deveci, M. (2023). A systematic review of trustworthy and explainable artificial intelligence in healthcare: Assessment of quality, bias risk, and data fusion. *Information Fusion*, *96*, 156-191. https://doi.org/10.1016/j.inffus.2023.03.008

Ali, M. M., Paul, B. K., Ahmed, K., Bui, F. M., Quinn, J. M. W., & Moni, M. A. (2021). Heart disease prediction using supervised machine learning algorithms: Performance analysis and comparison. *Computers in Biology and Medicine*, *136*, 104672. https://doi.org/10.1016/j.compbiomed.2021.104672

Alizadehsani, R., Roshanzamir, M., Abdar, M., Beykikhoshk, A., Khosravi, A., Panahiazar, M., Koohestani, A., Khozeimeh, F., Nahavandi, S., & Sarrafzadegan, N. (2019). A database for using machine learning and data mining techniques for coronary artery disease diagnosis. *Scientific Data*, *6*(1), 227. https://doi.org/10.1038/s41597-019-0206-3

Belete, D., & D H, M. (2021). Grid search in hyperparameter optimization of machine learning models for prediction of HIV/AIDS test results. *International Journal of Computers and Applications*, *44*, 1-12. https://doi.org/10.1080/1206212X.2021.1974663

Bradshaw, T. J., Huemann, Z., Hu, J., & Rahmim, A. (2023). A Guide to Cross-Validation for Artificial Intelligence in Medical Imaging. *Radiology: Artificial Intelligence*, *5*(4), e220232. https://doi.org/10.1148/ryai.220232

Charter of Fundamental Rights of the European Union, 202 OJ C (2016). http://data.europa.eu/eli/treaty/char_2016/oj/eng

Christopher M. Bishop. (2006). *Pattern Recognition and Machine Learning*.

Collins, G. S., Moons, K. G. M., Dhiman, P., Riley, R. D., Beam, A. L., Calster, B. V., Ghassemi, M., Liu, X., Reitsma, J. B., Smeden, M. van, Boulesteix, A.-L., Camaradou, J. C., Celi, L. A., Denaxas, S., Denniston, A. K., Glocker, B., Golub, R. M., Harvey, H., Heinze, G., … Logullo, P. (2024). TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ*, *385*, e078378. https://doi.org/10.1136/bmj-2023-078378

Deloitte. (2022). *Trustworthy Artificial Intelligence (AI)™*. Deloitte United States. https://www2.deloitte.com/us/en/pages/deloitte-analytics/solutions/ethics-of-ai-framework.html

Donders, A. R. T., Heijden, G. J. M. G. van der, Stijnen, T., & Moons, K. G. M. (2006). Review: A gentle introduction to imputation of missing values. *Journal of Clinical Epidemiology*, *59*(10), 1087-1091. https://doi.org/10.1016/j.jclinepi.2006.01.014

Elreedy, D., & Atiya, A. F. (2019). A Comprehensive Analysis of Synthetic Minority Oversampling Technique (SMOTE) for handling class imbalance. *Information Sciences*, *505*, 32-64. https://doi.org/10.1016/j.ins.2019.07.070

Endo, T., Watanabe, T., & Yamamoto, A. (2015). Confidence interval estimation by bootstrap method for uncertainty quantification using random sampling method. *Journal of Nuclear Science and Technology*, *52*(7-8), 993-999. https://doi.org/10.1080/00223131.2015.1034216

European Comission. (2019, abril 8). *Ethics guidelines for trustworthy AI*. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

European Comission. (2020, julio 17). *Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment*. https://digital-strategy.ec.europa.eu/en/library/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment

European Parliament. (2024). *Artificial Intelligence Act*.

García-Gómez, J. M., Tortajada, S., & Sáez, C. (2019). *Sistemas de Ayuda a la Decisión Médica*.

Huyen, C. (2022, febrero 7). *Data Distribution Shifts and Monitoring*. Chip Huyen. https://huyenchip.com/2022/02/07/data-distribution-shifts-and-monitoring.html

IBM. (2021, febrero 9). *Trustworthy AI*. IBM Research. https://research.ibm.com/topics/trustworthy-ai#our-work

John Clore, K. C. (2014). *Diabetes 130-US Hospitals for Years 1999-2008* [dataset]. [object Object]. https://doi.org/10.24432/C5230J

Kamiran, F., Mansha, S., Karim, A., & Zhang, X. (2018). Exploiting reject option in classification for social discrimination control. *Information Sciences*, *425*, 18-33. https://doi.org/10.1016/j.ins.2017.09.064

Karanam, S. (2021, agosto 11). *Curse of Dimensionality—A "Curse" to Machine Learning*. Medium. https://towardsdatascience.com/curse-of-dimensionality-a-curse-to-machine-learning-c122ee33bfeb

Krasanakis, E., Spyromitros-Xioufis, E., Papadopoulos, S., & Kompatsiaris, Y. (2018). Adaptive Sensitive Reweighting to Mitigate Bias in Fairness-aware Classification. *Proceedings of the 2018 World Wide Web Conference on World Wide Web - WWW '18*, 853-862. https://doi.org/10.1145/3178876.3186133

Molnar, C. (2021). *Aprendizaje automático interpretable*. https://fedefliguer.github.io/AAI/

Murthy, S., Abu Bakar, A., Abdul Rahim, F., & Ramli, R. (2019). A Comparative Study of Data Anonymization Techniques. *2019 IEEE 5th Intl Conference on Big Data Security on Cloud (BigDataSecurity), IEEE Intl Conference on High Performance and Smart Computing, (HPSC) and*

IEEE Intl Conference on Intelligent Data and Security (IDS), 306-309. https://doi.org/10.1109/BigDataSecurity-HPSC-IDS.2019.00063

Neupane, N. P. (2023, noviembre 3). Introduction to MLOps for Beginner—Part 1. *Medium*. https://netraneupane.medium.com/introduction-to-mlops-for-beginner-part-1-37207ea3004a

Ng, M. Y., Kapur, S., Blizinsky, K. D., & Hernandez-Boussard, T. (2022). The AI life cycle: A holistic approach to creating ethical AI for health decisions. *Nature medicine*, *28*(11), 2247-2249. https://doi.org/10.1038/s41591-022-01993-y

OHDSI. (2024). *Standardized Data: The OMOP Common Data Model*. https://www.ohdsi.org/data-standardization/

Olatunji, I. E., Rauch, J., Katzensteiner, M., & Khosla, M. (2022). A Review of Anonymization for Healthcare Data. *Big Data*, big.2021.0169. https://doi.org/10.1089/big.2021.0169

Ostling, S., Wyckoff, J., Ciarkowski, S. L., Pai, C.-W., Choe, H. M., Bahl, V., & Gianchandani, R. (2017). The relationship between diabetes mellitus and 30-day readmission rates. *Clinical Diabetes and Endocrinology*, *3*(1), 3. https://doi.org/10.1186/s40842-016-0040-x

Parlamento Europeo. (2016). *Reglamento general de protección de datos*.

R. Duda, P. Hart, & D. Stork. (2001). *Pattern Classification*.

Radley-Gardner, O., Beale, H., & Zimmermann, R. (Eds.). (2016). *Fundamental Texts On European Private Law*. Hart Publishing. https://doi.org/10.5040/9781782258674

Rančić, S., Radovanovic, S., & Delibašić, B. (2021). *Investigating Oversampling Techniques for Fair Machine Learning Models* (pp. 110-123). https://doi.org/10.1007/978-3-030-73976-8_9

Regenstrief Institute. (2024). *What LOINC is*. LOINC. https://loinc.org/get-started/what-loinc-is/

Regulation (EU) 2017/745 on Medical Devices, 117 OJ L (2017). http://data.europa.eu/eli/reg/2017/745/oj/eng

Sáez, C., Ferri, P., & García-Gómez, J. M. (2024). Resilient Artificial Intelligence in Health: Synthesis and Research Agenda Toward Next-Generation Trustworthy Clinical Decision Support. *Journal of Medical Internet Research*, *26*(1), e50295. https://doi.org/10.2196/50295

Sáez, C., Gutiérrez-Sacristán, A., Kohane, I., García-Gómez, J. M., & Avillach, P. (2020). EHRtemporalVariability: Delineating temporal data-set shifts in electronic health records. *GigaScience*, *9*(8), giaa079. https://doi.org/10.1093/gigascience/giaa079

Sáez, C., Martínez-Miranda, J., Robles, M., & García-Gómez, J. M. (2012). Organizing data quality assessment of shifting biomedical data. *Studies in Health Technology and Informatics*, *180*, 721-725.

Saltz, J. (2024, enero 31). The GenAI Life Cycle. *Data Science Process Alliance*. https://www.datascience-pm.com/the-genai-life-cycle/

Saltz, J. S. (2021). CRISP-DM for Data Science: Strengths, Weaknesses and Potential Next Steps. *2021 IEEE International Conference on Big Data (Big Data)*, 2337-2344. https://doi.org/10.1109/BigData52589.2021.9671634

Shang, Y., Jiang, K., Wang, L., Zhang, Z., Zhou, S., Liu, Y., Dong, J., & Wu, H. (2021). The 30-days hospital readmission risk in diabetic patients: Predictive modeling with machine learning classifiers. *BMC Medical Informatics and Decision Making*, *21*(Suppl 2), 57. https://doi.org/10.1186/s12911-021-01423-y

SPSS Modeler Subscription. (2021, agosto 17). https://www.ibm.com/docs/en/spss-modeler/saas?topic=dm-crisp-help-overview

Stavseth, M. R., Clausen, T., & Røislien, J. (2019). How handling missing data may impact conclusions: A comparison of six different imputation methods for categorical questionnaire data. *SAGE Open Medicine*, *7*, 2050312118822912. https://doi.org/10.1177/2050312118822912

Strack, B., DeShazo, J. P., Gennings, C., Olmo, J. L., Ventura, S., Cios, K. J., & Clore, J. N. (2014). Impact of HbA1c Measurement on Hospital Readmission Rates: Analysis of 70,000 Clinical Database

Patient Records. *BioMed Research International*, *2014*, e781670. https://doi.org/10.1155/2014/781670

Taskesen, E. (2023, abril 29). *Outlier Detection Using Principal Component Analysis and Hotelling's T2 and SPE/DmodX Methods*. Medium. https://towardsdatascience.com/outlier-detection-using-principal-component-analysis-and-hotellings-t2-and-spe-dmodx-methods-625b3c90897

Thiebes, S., Lins, S., & Sunyaev, A. (2021). Trustworthy artificial intelligence. *Electronic Markets*, *31*(2), 447-464. https://doi.org/10.1007/s12525-020-00441-4

UC Berkeley. (s. f.). *Universal Design Principles | Disability Access & Compliance*. Recuperado 23 de mayo de 2024, de https://dac.berkeley.edu/services/campus-building-accessibility/universal-design-principles

Understanding MLops Lifecycle: From Data to Deployment. (2024). ProjectPro. https://www.projectpro.io/article/mlops-lifecycle/885

United Nations. (2015). *Transforming Our World: The 2030 Agenda for Sustainable Development*. https://sdgs.un.org/publications/transforming-our-world-2030-agenda-sustainable-development-17981

Vilalta, R., Giraud-Carrier, C., & Brazdil, P. (2010). Meta-Learning—Concepts and Techniques. En O. Maimon & L. Rokach (Eds.), *Data Mining and Knowledge Discovery Handbook* (pp. 717-731). Springer US. https://doi.org/10.1007/978-0-387-09823-4_36

Vyhmeister, E., Castane, G., Östberg, P.-O., & Thevenin, S. (2023). A responsible AI framework: Pipeline contextualisation. *AI and Ethics*, *3*(1), 175-197. https://doi.org/10.1007/s43681-022-00154-8

Well-Architected machine learning lifecycle—Machine Learning Lens. (2023, julio). https://docs.aws.amazon.com/wellarchitected/latest/machine-learning-lens/well-architected-machine-learning-lifecycle.html

World Health Organization. (2024). *International Classification of Diseases (ICD)*. https://www.who.int/standards/classifications/classification-of-diseases

DOCUMENTO II

Presupuesto

Índice de tablas

Tabla 1. Cuadro de precios de mano de obra..... 86
Tabla 2. Cuadro de precios de maquinaria..... 86
Tabla 3. Cuadro de precios unitarios..... 87
Tabla 4. Cuadro de precios descompuestos..... 88
Tabla 5. Presupuestos parciales 90
Tabla 6. Presupuesto total de ejecución por contrata 91

1. Introducción

El siguiente documento tiene como objetivo estimar el coste total del proyecto relativo al Trabajo Fin de Grado (TFG) en Ingeniería Biomédica "Desarrollo de un sistema de Inteligencia Artificial Confiable en Salud". En el presupuesto se pretenden estimar los costes de mano de obra, materiales y maquinaria. Para realizar la estimación, partimos de una serie de suposiciones.

2. Cuadro de precios de mano de obra

La mano de obra requerida está formada por un Ingeniero Biomédico Junior, y el Tutor del TFG, Ingeniero Senior, encargado de supervisar su tarea. El coste para la empresa de cada trabajador viene dado por el salario bruto sumado a un coste adicional relativo a la Seguridad Social de entorno al 28%. Suponemos que el gasto para la empresa es de 22.000€ anuales para el Ingeniero Biomédico Estudiante y 65.000€ anuales para el Tutor del TFG. Con el fin de obtener el salario por hora, consideramos 366 días naturales, de los cuales descontamos sábados, domingos, festivos y vacaciones (24 días). Quedando un total de 228 días laborables con una jornada de 8 horas se tienen 1824 horas, lo que resulta un sueldo de 12,06 €/h para el estudiante y 35,63 €/h para el tutor. En la siguiente *Tabla 1* se muestra el cuadro de precios de mano de obra.

Tabla 1. *Cuadro de precios de mano de obra*

N°	Código	Designación	Importe		
			Precio (€)	Cantidad (Horas)	Total (€)
1	MO.IBE	Ingeniero Biomédico Junior	12,06	333,50	**4.022,01**
2	MO.TUT	Ingeniero Tutor del TFG	35,63	57,30	**2.041,35**
				Total mano de obra	**6.063,36**

3. Cuadro de precios de maquinaria

El cuadro de precios de maquinaria muestra de forma aproximada el coste que supone toda la maquinaria empleada en la elaboración del proyecto. Debido a que los softwares de programación incluidos son en su gran mayoría gratuitos, tan solo contamos con los costes del ordenador portátil propio del estudiante y los programas asociados a su uso, Licencia Windows 10 y Office 365. En la *Tabla 2* apreciamos el resultado:

Tabla 2. *Cuadro de precios de maquinaria*

N°	Código	Designación	Importe		
			Precio (€)	Cantidad	Total (€)
1	MAQ.LO365	Licencia de Microsoft Office 365	0,90	90,00 h	**81,00**
2	MAQ.LW11	Licencia Windows 11	0,90	311,00 h	**279,90**
3	MAQ.OPP	Ordenador Portatil Personal	1,50	321,00 h	**481,50**
				Total Maquinaria	**842,40**

4. Cuadro de precios unitarios

El cuadro de precios unitarios mostrado en la *Tabla 3* describe cada unidad y su respectivo importe. Los precios unitarios provienen de los descompuestos, cuya descomposición se realizará en el apartado 5.

Tabla 3. *Cuadro de precios unitarios*

N°	Designación	Importe	
		En cifra (Euros)	En letra (Euros)
	01 Definición del proyecto		
01.01	h Reunión inicial con el tutor del TFG	**49,60 €**	CUARENTA Y NUEVE EUROS CON SESENTA CÉNTIMOS
01.02	h Reunión de planificación del trabajo	**49,60 €**	CUARENTA Y NUEVE EUROS CON SESENTA CÉNTIMOS
	02 Investigación del estado del arte		
02.01	h Revisión del estado del arte	**15,97 €**	QUINCE EUROS CON NOVENTA Y SIETE CÉNTIMOS
02.02	h Búsqueda de métodos técnicos viables	**15,04 €**	QUINCE EUROS CON CUATRO CÉNTIMOS
02.03	h Búsqueda de conjuntos de datos viables	**24,30 €**	VEINTICUATRO EUROS CON TREINTA CÉNTIMOS
	03 Recopilación y tratamiento de los datos		
03.01	h Búsqueda de conjuntos de datos viables	**33,57 €**	TREINTA Y TRES EUROS CON CINCUENTA Y SIETE CÉNTIMOS
03.02	h Caracterización y acondicionamiento de los datos	**24,30 €**	VEINTICUATRO EUROS CON TREINTA CÉNTIMOS
	04 Desarrollo del pipeline semiautomático		
04.01	h Definición de la matriz de requisitos según ciclo de vida	**52,09 €**	CINCUENTA Y DOS EUROS CON NUEVE CÉNTIMOS
04.02	h Desarrollo de las libretas de código genéricas	**18,74 €**	DIECIOCHO EUROS CON SETENTA Y CUATRO CÉNTIMOS
04.03	h Evaluación e interpretación de los resultados con diferentes conjuntos de datos	**24,30 €**	VEINTICUATRO EUROS CON TREINTA CÉNTIMOS
	05 Redacción y defensa del TFG		
05.01	h Redacción de los documentos	**15,97 €**	QUINCE EUROS CON NOVENTA Y SIETE CÉNTIMOS
05.02	h Revisión de los documentos y corrección de errores	**34,78 €**	TREINTA Y CUATRO EUROS CON SETENTA Y OCHO CÉNTIMOS
05.03	h Preparación exposición	**51,16 €**	CINCUENTA Y UN EUROS CON DIECISEIS CÉNTIMOS

5. Cuadro de precios descompuestos

En la *Tabla 4* podemos observar el cuadro de precios descompuestos, cuyo fin es detallar los distintos capítulos según las unidades de obra y maquinaria, así como los tiempos requeridos para cada una.

Tabla 4. Cuadro de precios descompuestos

01 Definición del proyecto

Código	Ud	Descripción				Total
01.01	h	Reunión inicial con el tutor del TFG				
1,00	h	Ingeniero Biomédico Junior			12,06 €	12,06 €
1,00	h	Ingeniero Tutor del TFG			35,63 €	35,63 €
			4,00 %	Costes indirectos	47,69 €	**1,91 €**
				Precio total por h		**49,60 €**
01.02	h	Reunión de planificación del trabajo				
1,00	h	Ingeniero Biomédico Junior			12,06 €	12,06 €
1,00	h	Ingeniero Tutor del TFG			35,63 €	35,63 €
			4,00 %	Costes indirectos	47,69 €	**1,91 €**
				Precio total por h		**49,60 €**

02 Investigación del estado del arte

Código	Ud	Descripción				Total
02.01	h	Revisión del estado del arte				
1,00	h	Ingeniero Biomédico Junior			12,06 €	12,06 €
1,00	h	Licencia de Microsoft Office 365			0,90 €	0,90 €
1,00	h	Ordenador Portatil Personal			1,50 €	1,50 €
1,00	h	Licencia Windows 11			0,90 €	0,90 €
			4,00 %	Costes indirectos	15,36 €	**0,61 €**
				Precio total por h		**15,97 €**
02.02	h	Búsqueda de métodos técnicos viables				
1,00	h	Ingeniero Biomédico Junior			12,06 €	12,06 €
1,00	h	Ordenador Portatil Personal			1,50 €	1,50 €
1,00	h	Licencia Windows 11			0,90 €	0,90 €
			4,00 %	Costes indirectos	14,46 €	**0,58 €**
				Precio total por h		**15,04 €**
02.03	h	Búsqueda de conjuntos de datos viables				
1,00	h	Ingeniero Biomédico Junior			12,06 €	12,06 €
0,25	h	Ingeniero Tutor del TFG			35,63 €	8,91 €
1,00	h	Ordenador Portatil Personal			1,50 €	1,50 €
1,00	h	Licencia Windows 11			0,90 €	0,90 €
			4,00 %	Costes indirectos	23,37 €	**0,93 €**
				Precio total por h		**24,30 €**

03 Recopilación y tratamiento de los datos

Código	Ud	Descripción				Total
03.01	h	Búsqueda de conjuntos de datos viables				
1,00	h	Ingeniero Biomédico Junior			12,06 €	12,06 €
0,50	h	Ingeniero Tutor del TFG			35,63 €	17,82 €
1,00	h	Ordenador Portatil Personal			1,50 €	1,50 €
1,00	h	Licencia Windows 11			0,90 €	0,90 €
			4,00 %	Costes indirectos	32,28 €	**1,29 €**
				Precio total por h		**33,57 €**
03.02	h	Caracterización y acondicionamiento de los datos				
1,00	h	Ingeniero Biomédico Junior			12,06 €	12,06 €
0,25	h	Ingeniero Tutor del TFG			35,63 €	8,91 €
1,00	h	Licencia Windows 11			0,90 €	0,90 €
1,00	h	Ordenador Portatil Personal			1,50 €	1,50 €
			4,00 %	Costes indirectos	23,37 €	**0,93 €**
				Precio total por h		**24,30 €**

04 Desarrollo del pipeline semiautomático

Código	Ud	Descripción				Total
04.01	**h**	Definición de la matriz de requisitos según ciclo de vida				
1,00	h	Ingeniero Biomédico Junior			12,06 €	12,06 €
1,00	h	Ingeniero Tutor del TFG			35,63 €	35,63 €
1,00	h	Ordenador Portatil Personal			1,50 €	1,50 €
1,00	h	Licencia Windows 11			0,90 €	0,90 €
			4,00	% Costes indirectos	50,09 €	**2,00 €**
				Precio total por h		**52,09 €**
04.02	**h**	Desarrollo de las libretas de código genéricas				
1,00	h	Ingeniero Biomédico Junior			12,06 €	12,06 €
0,10	h	Ingeniero Tutor del TFG			35,63 €	3,56 €
1,00	h	Ordenador Portatil Personal			1,50 €	1,50 €
1,00	h	Licencia Windows 11			0,90 €	0,90 €
			4,00	% Costes indirectos	18,02 €	**0,72 €**
				Precio total por h		**18,74 €**
04.03	**h**	Evaluación e interpretación de los resultados con diferentes conjuntos de datos				
1,00	h	Ingeniero Biomédico Junior			12,06 €	12,06 €
0,25	h	Ingeniero Tutor del TFG			35,63 €	8,91 €
1,00	h	Ordenador Portatil Personal			1,50 €	1,50 €
1,00	h	Licencia Windows 11			0,90 €	0,90 €
			4,00	% Costes indirectos	23,37 €	**0,93 €**
				Precio total por h		**24,30 €**

05 Redacción y defensa del TFG

Código	Ud	Descripción				Total
05.01	**h**	Redacción de los documentos				
1,00	h	Ingeniero Biomédico Junior			12,06 €	12,06 €
1,00	h	Ordenador Portatil Personal			1,50 €	1,50 €
1,00	h	Licencia Windows 11			0,90 €	0,90 €
1,00	h	Licencia de Microsoft Office 365			0,90 €	0,90 €
			4,00	% Costes indirectos	15,36 €	**0,61 €**
				Precio total por h		**15,97 €**
05.02	**h**	Revisión de los documentos y corrección de errores				
1,00	h	Ingeniero Biomédico Junior			12,06 €	12,06 €
0,60	h	Ingeniero Tutor del TFG			35,63 €	21,38 €
			4,00	% Costes indirectos	33,44 €	**1,34 €**
				Precio total por h		**34,78 €**
05.03	**h**	Preparación exposición				
1,00	h	Ingeniero Biomédico Junior			12,06 €	12,06 €
1,00	h	Ingeniero Tutor del TFG			35,63 €	35,63 €
1,00	h	Ordenador Portatil Personal			1,50 €	1,50 €
			4,00	% Costes indirectos	49,19 €	**1,97 €**
				Precio total por h		**51,16 €**

6. Presupuestos parciales

En la *Tabla 5* se muestran tanto los presupuestos parciales como las mediciones. En la tabla se define cada unidad con su respectiva cantidad e importe para cada uno de los distintos capítulos.

Tabla 5. Presupuestos parciales

Presupuesto parcial nº 01 Definición del proyecto

Nº	Ud	Descripción		Medición	Precio	Importe
01.01	H	Reunión inicial con el tutor del TFG				
			Total h :	2,00	49,60	**99,20**
01.02	H	Reunión de planificación del trabajo				
			Total h :	2,50	49,60	**124,00**
		Total Presupuesto parcial nº 01 Definición del proyecto :				**223,20**

Presupuesto parcial nº 02 Investigación del estado del arte

Nº	Ud	Descripción		Medición	Precio	Importe
02.01	H	Revisión del estado del arte				
			Total h :	30,00	15,97	**479,10**
02.02	H	Búsqueda de métodos técnicos viables				
			Total h :	2,00	15,04	**30,08**
02.03	H	Búsqueda de conjuntos de datos viables				
			Total h :	20,00	24,30	**486,00**
		Total Presupuesto parcial nº 02 Investigación del estado del arte :				**995,18**

Presupuesto parcial nº 03 Recopilación y tratamiento de los datos

Nº	Ud	Descripción		Medición	Precio	Importe
03.01	H	Búsqueda de conjuntos de datos viables				
			Total h :	8,00	33,57	**268,56**
03.02	H	Caracterización y acondicionamiento de los datos				
			Total h :	6,00	24,30	**145,80**
		Total Presupuesto parcial nº 03 Recopilación y tratamiento de los datos :				**414,36**

Presupuesto parcial nº 04 Desarrollo del pipeline semiautomático

Nº	Ud	Descripción		Medición	Precio	Importe
04.01	H	Definición de la matriz de requisitos según ciclo de vida				
			Total h :	5,00	52,09	**260,45**
04.02	H	Desarrollo de las libretas de código genéricas				
			Total h :	150,00	18,74	**2.811,00**
04.03	H	Evaluación e interpretación de los resultados con diferentes conjuntos de datos				
			Total h :	30,00	24,30	**729,00**
		Total Presupuesto parcial nº 04 Desarrollo del pipeline semiautomático :				**3.800,45**

Presupuesto parcial nº 05 Redacción y defensa del TFG

Nº	Ud	Descripción	Medición	Precio	Importe
05.01	H	Redacción de los documentos			
		Total h :	60,00	15,97	**958,20**
05.02	H	Revisión de los documentos y corrección de errores			
		Total h :	8,00	34,78	**278,24**
05.03	H	Preparación exposición			
		Total h :	10,00	51,16	**511,60**
		Total Presupuesto parcial nº 05 Redacción y defensa del TFG :			**1.748,04**

7. Presupuesto total de ejecución por contrata

Para finalizar, a partir de los Presupuestos Parciales individualizados por capítulos, obtenemos el Presupuesto total de Ejecución Material. Sumando a esta cifra se le añade un 16% de gastos generales y un 7% de beneficio industrial. Esta cantidad será la base imponible a la que aplicar el 21% de IVA impuesto por la Agencia Tributaria para obtener finalmente el Presupuesto de Ejecución por Contrata. En la *Tabla 6* observamos el resultado de dichos cálculos.

Tabla 6. Presupuesto total de ejecución por contrata

Proyecto: Presupuesto TFG

Capítulo	Importe
1 Definición del proyecto	223,20
2 Investigación del estado del arte	995,18
3 Recopilación y tratamiento de los datos	414,36
4 Desarrollo del pipeline semiautomático	3.800,45
5 Redacción y defensa del TFG	1.748,04
Presupuesto de ejecución material	7.181,23
16% de gastos generales	1.149,00
7% de beneficio industrial	502,69
Suma	8.832,92
21% IVA	1.854,91
Presupuesto de ejecución por contrata	10.687,83

Asciende el presupuesto de ejecución por contrata a la expresada cantidad de DIEZ MIL SEISCIENTOS OCHENTA Y SIETE EUROS CON OCHENTA Y TRES CÉNTIMOS.

Valencia, Junio 2024
Ingeniero Biomédico

Carlos de Manuel Vicente