Universidad Politécnica
de Madrid

Escuela Técnica Superior de Ingenieros Informáticos

Máster Universitario en Inteligencia Artificial

Trabajo Fin de Máster

Detección de la Fragilidad en Personas Mayores mediante Modelos de Inteligencia Artificial: Un Enfoque Explicable e Integrado en la Práctica Clínica

Autor(a): Adrián Arana Hernández
Tutor(a): Esteban García Cuesta y Elena Villalba Mora

Madrid, Julio - 2024

Este Trabajo Fin de Máster se ha depositado en la ETSI Informáticos de la Universidad Politécnica de Madrid para su defensa.

Trabajo Fin de Máster
Máster Universitario en Inteligencia Artificial

Título: Detección de la Fragilidad en Personas Mayores mediante Modelos de Inteligencia Artificial: Un Enfoque Explicable e Integrado en la Práctica Clínica

Julio - 2024

Autor(a): Adrián Arana Hernández
Tutor(a): Esteban García Cuesta
Inteligencia Artificial
ETSI Informáticos
Universidad Politécnica de Madrid
Tutor(a): Elena Villalba Mora
Lenguajes y Sistemas Informáticos e Ingeniería de Software
ETSI Informáticos
Universidad Politécnica de Madrid

Resumen

La fragilidad en personas mayores es un problema de salud significativo que puede llevar a una disminución en la calidad de vida, aumentar el riesgo de caídas, hospitalizaciones y mortalidad, y que además será un problema más común en unos años. La motivación detrás de este estudio es encontrar formas más eficaces y accesibles de evaluar y predecir la fragilidad para mejorar la intervención temprana y el manejo clínico de estas personas. El estudio se centra en la creación de modelos de inteligencia artificial que predicen la fragilidad utilizando el test FTS-5 como referencia.

El test FTS-5, aunque efectivo, puede ser complicado de aplicar en centros no especializados debido a la complejidad de los ítems requeridos. Por ello, el objetivo del estudio es desarrollar modelos que mantengan una alta precisión diagnóstica, pero que requieran variables más simples y de fácil obtención. Esto se hace para facilitar su implementación en entornos clínicos con recursos limitados y mejorar la accesibilidad del diagnóstico de fragilidad.

El uso de la inteligencia artificial se justifica por su capacidad de analizar grandes volúmenes de datos y detectar patrones complejos que no son evidentes mediante métodos tradicionales. La justificación detrás del empleo de ítems más sencillos, manteniendo la mayor precisión posible, no solo haría el test más accesible, sino que también permitiría un diagnóstico más rápido y eficiente, mejorando así la intervención temprana y la gestión del cuidado de los pacientes, además de facilitar el trabajo de los profesionales sanitarios.

En términos de metodología, el estudio emplea diversos algoritmos de aprendizaje automático, como Support Vector Machine, k-Nearest Neighbour, Random Forest, Ada Boost y XGBoost, aplicados a conjuntos de datos obtenidos del Estudio Toledo de Envejecimiento Saludable. En un segundo paso, se emplean técnicas de reducción de la dimensionalidad y se exploran métodos para entender el comportamiento de estos modelos.

Los resultados iniciales muestran que es posible utilizar ítems más simples que los del test FTS-5 comprometiendo mínimamente la precisión diagnóstica. Estos hallazgos sugieren que los modelos de inteligencia artificial pueden simplificar las evaluaciones de fragilidad y hacerlas más accesibles, abriendo la puerta a su aplicación en diversos entornos clínicos, especialmente aquellos con recursos limitados.

Abstract

Frailty in the elderly is a significant health problem that can lead to decreased quality of life, increased risk of falls, hospitalizations, and mortality, and is expected to become more common in the coming years. The motivation behind this study is to find more effective and accessible ways to assess and predict frailty to improve early intervention and clinical management. The study focuses on creating artificial intelligence models that predict frailty using the FTS-5 test as a reference.

The FTS-5 test, although effective, can be complicated to apply in non-specialized centers due to the complexity of the required items. Therefore, the study aims to develop models that maintain high diagnostic accuracy but require simpler and easier-to-obtain variables. This is done to facilitate implementation in clinical settings with limited resources and to improve the accessibility of frailty diagnosis.

The use of artificial intelligence is justified by its ability to analyze large volumes of data and detect complex patterns that are not evident through traditional methods. The rationale behind using simpler items while maintaining the highest possible accuracy would not only make the test more accessible but also allow for faster and more efficient diagnosis, thereby improving early intervention and patient care management, as well as easing the workload of healthcare professionals.

In terms of methodology, the study employs various machine learning algorithms, such as Support Vector Machine, k-Nearest Neighbour, Random Forest, Ada Boost, and XGBoost, applied to datasets obtained from the Toledo Study for Healthy Aging. In a second step, dimensionality reduction techniques are employed, and methods to understand the behavior of these models are explored.

Initial results show that it is possible to use simpler items than those in the FTS-5 test while minimally compromising diagnostic accuracy. These findings suggest that artificial intelligence models can simplify frailty assessments and make them more accessible, opening the door to their application in various clinical settings, especially those with limited resources.

Tabla de contenidos

Capítulo 1

Introducción

La fragilidad es un síndrome geriátrico que afecta a una gran parte de la población en las personas mayores de 65 años. Este síndrome se caracteriza por una disminución de la reserva fisiológica. Además, se traduce directamente en un claro aumento del riesgo de incapacidad para la persona, pérdida de resistencia y una mayor vulnerabilidad [9]. El aumento de la esperanza de vida en el siglo XXI es una de las causas más evidentes de la aparición del síndrome de fragilidad, y es que, se estima que en el año 2050 las personas mayores de 60 años ocuparán en torno al 22 % de la población mundial [25]. Esto implica que el síndrome de la fragilidad vaya a convertirse en uno de los principales problemas para el ámbito sanitario, tanto a nivel de salud como económico, al requerir más costos para poder atender el creciente número de pacientes.

En este contexto, surge la necesidad de desarrollar métodos eficaces y accesibles para evaluar y predecir la fragilidad en personas mayores. La utilización de modelos de inteligencia artificial se presenta como una solución prometedora, permitiendo analizar grandes volúmenes de datos y detectar patrones complejos que no son evidentes mediante métodos tradicionales.

Este estudio se centra en la creación de modelos predictivos basados en el test FTS-5 [8], una escala eficaz y ampliamente utilizada para evaluar la fragilidad.

El test FTS-5, aunque efectivo, puede ser complicado de aplicar en centros no especializados debido a la complejidad de los ítems requeridos. Por ello, el objetivo del estudio es desarrollar modelos que mantengan una alta precisión diagnóstica, pero que requieran variables más simples y de fácil obtención. Esto se hace para facilitar su implementación en entornos clínicos con recursos limitados y mejorar la accesibilidad del diagnóstico de fragilidad e incluso poder realizar una detección precoz mediante cuestionarios telefónicos de corta duración.

Por ello, los objetivos principales que se plantean a lo largo del proyecto son el análisis exhaustivo de los datos a utilizar, incluyendo su procesamiento, la búsqueda de factores determinantes para la fragilidad y la creación de modelos que permitan una detección de la fragilidad con menos recursos que la escala objetivo, la FTS-5.

Capítulo 2

Estado del Arte

En este capítulo se va a explorar la literatura relacionada con la fragilidad, con técnicas de inteligencia artificial, en especial sobre explicabilidad para la detección de características relevantes y, además, se trata de buscar algunas aplicaciones similares a la que se presenta en este proyecto.

2.1. Fragilidad

En la literatura existen diversos tests para evaluar la fragilidad en las personas mayores. Algunos de los más conocidos son el test Linda Fried [5] o el test de Rockwood [20] y, pese a presentar algunas ventajas a la hora de medir la fragilidad, también cuentan con algunos inconvenientes respecto a otro tipo de tests o escalas.

2.1.1. Linda Fried

Este test consta de 5 ítems:

1. Pérdida de peso no intencional: se refiere a una pérdida de peso corporal sin haber hecho cambios significativos en la dieta o en la actividad diaria que puedan justificar esta pérdida.
2. Debilidad en la fuerza del agarre: se mide utilizando un dinamómetro para evaluar la fuerza de los músculos de la mano. Una menor fuerza de agarre es un signo de debilidad muscular, la cual suele estar relacionada con la fragilidad.
3. Sensación de agotamiento: percepción subjetiva de estar continuamente cansado y sin energía. Se suele evaluar mediante cuestionarios.
4. Velocidad de la marcha: es la velocidad al caminar. Se mide en una distancia corta y es un indicador de la movilidad y la capacidad funcional de la persona. Una velocidad de marcha reducida sugiere deterioro en la salud.
5. Bajo nivel de actividad física: cantidad de actividad física que realiza la persona en su vida diaria. Se suele medir a través de cuestionarios.

Si una persona cumple con al menos 3 de los 5 ítems se considerará frágil [5].

La principal ventaja de este test es el empleo de pruebas fáciles de aplicar, por norma general. Debido a esto, se puede afirmar que es un buen test para usar en la detección

precoz de la fragilidad.

La principal desventaja de este test es el uso de preguntas subjetivas, que tienen la misma importancia que las pruebas objetivas, en las que el paciente puede no responder adecuadamente o, incluso, al ser subjetivas, pueden ser valoradas de diferente forma para distintos pacientes.

2.1.2. Rockwood

Esta escala también es conocida como la Escala de Fragilidad Clínica (CSF). Para evaluar a un paciente se consideran distintos aspectos como:

1. Nivel de actividad física: evalúa cuánto ejercicio o actividad física realiza la persona regularmente. Se considera tanto la actividad moderada como la intensa.

2. Enfermedades crónicas: se refiere a la presencia y el número de enfermedades crónicas que tiene la persona.

3. Independencia en actividades de la vida diaria: mide la capacidad de la persona para realizar actividades diarias esenciales de forma independiente, como vestirse o bañarse.

4. Energía y Vitalidad: evalúa la sensación general de energía y vitalidad de la persona. Se suele medir mediante cuestionarios.

5. Estado funcional: valora la capacidad de la persona para realizar actividades cotidianas más complejas, como manejar finanzas.

6. Estado cognitivo: examina las funciones cognitivas de la persona, como la memoria, el pensamiento crítico o la capacidad de toma de decisiones.

7. Pronóstico de salud: es la perspectiva general de salud de la persona en el futuro, considerando su estado actual de salud, la presencia de enfermedades crónicas u otros factores que afecten directamente a la calidad de vida.

Esta escala no solo clasifica a los pacientes como frágiles o no frágiles, sino que los clasifica en distintas categorías [20].

La principal ventaja de esta escala es su precisión, ya que al asignar a los pacientes en distintas categorías se da una clasificación más exacta del estado de los mismos. Además, se tiene una visión más global al incorporar muchas pruebas que evalúan aspectos muy distintos. Como principal desventaja, se muestra su complejidad, al requerir de tantas pruebas y algunas complejas, no es muy apropiada como técnica de evaluación si no es en un centro especializado.

2.1.3. Frailty Trait Scale y FTS-5

A continuación, se van a mostrar otros test desarrollados más recientemente para medir la fragilidad en los pacientes y los cuales han demostrado ser válidos. Estos test son el Frailty Trait Scale (FTS) [7] y el Frailty Trait Scale 5 (FTS-5) [8]. Es importante destacar que el FTS-5 es una adaptación del FTS en la que se utilizan menos ítems.

2.1.3.1. Frailty Trait Scale

Esta escala es una innovadora forma de medir el estado de fragilidad de un paciente. El objetivo principal de esta escala es ser capaz de realizar una clasificación de la fragilidad que capture las distintas dimensiones que tiene por naturaleza este síndrome, ya que muchas otras escalas tienen un enfoque más centrado en ciertos aspectos como puede ser el estado físico o el número de enfermedades crónicas del sujeto.

Este test se compone de varios ítems que recogen distintas dimensiones de la fragilidad, como pueden ser la función física, el estado nutricional, la función cognitiva, la salud psicológica, el apoyo social o la comorbilidad. Estas dimensiones son:

1. **Balance energético y nutrición**

   a) Índice de Masa Corporal (IMC): se calcula dividiendo el peso de una persona en kilogramos por su altura en metros al cuadrado. Un IMC fuera del rango normal puede indicar problemas de salud relacionados con el peso.

   b) Obesidad central (circunferencia de cintura): es un indicador de la cantidad de grasa abdominal, que está asociada con un mayor riesgo de enfermedades cardiovasculares y metabólicas.

   c) Nivel de albúmina sérica: es una proteína producida por el hígado, y sus niveles en sangre pueden indicar el estado nutricional de una persona.

   d) Pérdida de peso involuntaria: se refiere a una pérdida de peso corporal sin haber hecho cambios significativos en la dieta o en la actividad diaria que puedan justificar esta pérdida.

2. **Actividad física**

   a) Total de actividad física: evaluado mediante la puntuación total de la Escala de Actividad Física para Ancianos (PASE), que mide la cantidad y el tipo de actividad física realizada por una persona mayor en su vida diaria.

3. **Sistema nervioso**

   a) Fluidez verbal: mide la cantidad de nombres de animales que una persona puede enumerar en un minuto, lo cual es un indicador de la función cognitiva.

   b) Balance: evaluado mediante el test de Romberg, que mide la capacidad de una persona para mantener el equilibrio de pie con los ojos cerrados.

4. **Sistema vascular**

   a) Índice Braquial-Tobillo (ABI) medido con ultrasonido Doppler: compara la presión sanguínea en el tobillo con la del brazo para detectar enfermedades arteriales periféricas.

5. **Debilidad**

   a) Fuerza de prensión manual: se mide utilizando un dinamómetro para evaluar la fuerza de los músculos de la mano. Una menor fuerza de agarre es un signo de debilidad muscular, la cual suele estar relacionada con la fragilidad.

b) Fuerza de extensión de rodilla: medida con un dinamómetro Lafayette, esta prueba evalúa la fuerza de los músculos de la pierna.

6. **Baja energía**

   a) Test de levantadas: mide el número de veces que una persona puede levantarse de una silla en 30 segundos, evaluando así la fuerza y resistencia muscular de las piernas.

7. **Lentitud**

   a) Velocidad de la marcha: mide el tiempo necesario para caminar 3 metros a un ritmo normal. Este es un indicador de la salud y la capacidad funcional.

Cada ítem se puntúa de 0 a 4, siendo el 0 un mejor desempeño y 4 el peor, a excepción del test de levantadas que va de 0 a 5, siendo nuevamente 0 el mejor desempeño y 5 el peor. Los ítems se analizan según la distribución en quintiles de la población estudiada. La puntuación final se calcula sumando las puntuaciones de cada ítem y calculando el porcentaje de puntos posibles para que la puntuación máxima sea 100. Los sujetos que alcancen una puntuación de 50 o más se considerarán frágiles [7].

La FTS tiene una serie de ventajas con respecto a otras herramientas empleadas en la literatura para medir la fragilidad, por ejemplo:

- **Multidimensionalidad:** la FTS es una escala que abarca un gran número de aspectos distintos de los sujetos, recogiendo en 12 ítems, los 7 aspectos vistos anteriormente a analizar.
- **Validación:** esta escala ha demostrado ser capaz de predecir resultados adversos de salud, como hospitalización o caídas.

2.1.3.2. FTS-5

Tras haber analizado la escala FTS surge un problema principal, hace falta registrar 12 ítems para valorar el estado de fragilidad de un paciente, lo que se puede traducir en una alta complejidad de evaluación comparado con otras escalas, pese a analizar más dimensiones. Esto complica considerablemente que pueda ser usado como medida de detección precoz, por ello aparece la Frailty Trait Scale de 5 ítems (FTS-5).

Los ítems que mejor optimizaron la capacidad de predicción fueron:

1. Índice de Masa Corporal (IMC): como se explicó en apartados anteriores, se calcula dividiendo el peso de una persona en kilogramos por su altura en metros al cuadrado. Este índice se utiliza para clasificar el peso de una persona y detectar posibles problemas de salud relacionados con el peso.
2. Test de Romberg progresivo: evalúa el equilibrio de una persona. El paciente debe mantenerse en bipedestación con los pies juntos y los ojos cerrados. La progresión del test implica aumentar la dificultad, por ejemplo con una sola pierna.
3. Actividad física medida con la Escala de Actividad Física para Ancianos (PASE): es una escala que mide la actividad física de los ancianos en sus actividades diarias, incluyendo tareas domésticas, ejercicios y actividades recreativas. Se basa en un cuestionario que evalúa la cantidad y frecuencia de estas actividades en una semana.

4. Velocidad de la marcha: mide la rapidez con la que una persona puede caminar una distancia específica, generalmente de 3 a 10 metros. Se utiliza para evaluar la movilidad y la función física general de la persona.
5. Fuerza de prensión manual: se mide utilizando un dinamómetro, que evalúa la fuerza de agarre de la mano. Este test es un indicador de la fuerza muscular general.

Cada uno de estos ítems podría obtener una puntuación del 0 al 10, a excepción del test de Romberg progresivo, el cual podría tener las puntuaciones 0, 2.5, 5, 7.5 y 10. En todas las pruebas una puntuación mayor significa un peor desempeño. En la Figura 2.1 se pueden observar los rangos y puntuaciones de cada prueba, destacando además, que hay distinciones según el género del sujeto tanto en el IMC como en la fuerza de prensión manual.

FTS Short Forms Scoring Table

Score	BMI, kg/m^2		PASE	Gait Speed, s*	Grip Strength, kg		Score	Progressive Romberg	
					Women	Men		Position	Seconds
0	23.01-26.99		>194	<2.45	>22	>29	0	Tandem	≥10
1	27-28.99	21.01-23	174.61-194	2.45-2.99	19.81-22	26.11-29	2.5	Tandem	3.01-9.99
2	29-30.99	19.01-21	155.21-174.6	3.00-3.54	17.61-19.8	23.21-26.1	5	Tandem	≤3
3	31-32.99	17.01-19	135.81-155.2	3.55-4.09	15.41-17.6	20.31-23.2		Semitandem	≥10
4	33-34.99	15.01-17	116.41-135.8	4.10-4.64	13.21-15.4	17.41-20.3	7.5	Semitandem	<10
5	35-36.99	13.01-15†	97.01-116.4	4.65-5.19	11.01-13.2	14.51-17.4		Side by side	≥10
6	37-38.99	11.01-13†	77.61-97	5.20-5.74	8.81-11.0	11.61-14.5	10	Side by side	<10
7	39-40.99	NA	58.21-77.6	5.75-6.29	6.61-8.8	8.71-11.6			
8	41-42.99	NA	38.81-58.2	6.30-6.84	4.41-6.6	5.81-8.7			
9	43-44.99	NA	19.41-38.8	6.85-7.39	2.21-4.4	2.91-5.8			
10	≥45	NA	0-19.4	≥7.4	0-2.2	0-2.9			

PASE, Physical Activity Scale for the Elderly.
FTS_5 includes all the items of the table (range 0-50), and frail participants are those with FTS_5 scores >25. FTS_3 includes BMI, PASE, and Romberg test (range 0-30), and frail participants are those with FTS_3 scores >15.
*Gait speed refers to time in accomplish 3-metres at usual pace.
†Model estimation.

Figura 2.1: Puntuaciones de las escalas FTS-5 y FTS-3 [8].

Por otra parte, en el estudio de la FTS-5 se concluye que tiene un área bajo la curva (AUC) más alta en ciertos eventos de salud que otras escalas de fragilidad, superando incluso a la FTS, aunque sin diferencias estadísticamente significativas.

El punto de corte para calificar a una persona como frágil se sitúa en puntuaciones mayores o iguales que 25 en la suma total de las pruebas [8].

Las principales ventajas que presenta la FTS-5 son:

- **Multidimensionalidad:** pese a tener menos ítems que la FTS, sigue abarcando distintas dimensiones del sujeto, traduciéndose en una escala que tiene en cuenta factores muy diversos.
- **Validación:** esta escala ha sido capaz de obtener resultados bastante buenos, llegando a superar en ciertos aspectos a su forma original, la FTS.
- **Flexibilidad:** al tener pocos ítems, siendo solo 5, se facilita la evaluación de fragilidad a los sujetos, por lo que se puede considerar una herramienta para la evaluación precoz.

2.2. Explicabilidad

En el aprendizaje de modelos de inteligencia artificial, en ocasiones, es de vital importancia conocer cómo estos toman decisiones, en especial en ámbitos relacionados con la medicina, como en este caso.

Por esto, el que un modelo sea capaz de dar al usuario una explicabilidad se vuelve un factor crucial a la hora de buscar factores determinantes para el desarrollo de una enfermedad.

Es en este caso cuando la inteligencia artificial explicable adquiere un rol determinante mediante la aplicación de técnicas que pueden dar mayor valor a un modelo de inteligencia artificial e incluso quitárselo, al detectar modelos no realistas o cuya compatibilidad con el razonamiento humano es imposible.

Algunas de estas técnicas más novedosas y avanzadas se muestran en las siguientes subsecciones.

2.2.1. Shapley Additive Explanations (SHAP)

Este método está basado en la teoría de juegos y los valores Shapley [22]. Estos valores son una forma de distribuir de manera justa las ganancias que genera una coalición, la cual será, en modelos de machine learning, un conjunto de características [11].

2.2.2. Local Interpretable Model-Agnostic Explanations (LIME)

Está técnica se utilizar para aportar explicabilidad a las predicciones de un modelo de tipo caja-negra de forma local. Este método, además, es agnóstico de modelo, es decir, puede aplicarse a cualquier tipo de modelo [19].

2.3. Aplicaciones similares

En el ámbito de la medicina son numerosas las aplicaciones en las que se emplea la inteligencia artificial para la detección de alguna enfermedad. Algunas en las que se presenta un escenario similar al de este proyecto y por lo tanto sirven de inspiración son:

- **Clasificadoras de ML para la predicción del cáncer de cuello uterino**: en este caso, se presenta un ejemplo de aplicación en el que se destaca el empleo de la técnica Stratified K-fold cross-validation (SKCV) [21], y que además, presenta un escenario muy similar al de este proyecto, con clases desbalanceadas y donde se le da más importancia a métricas como la *precision* y el *recall*, que a otras como el *accuracy* [17].
- **Clasificadores para la clasificación del cáncer de mama**: al igual que en la aplicación anterior, se presenta otra aplicación de técnicas de machine learning para la detección de enfermedades, además de la misma manera emplea SKCV y tiene un desbalanceo de clases considerable [13].

- **Diagnóstico de la enfermedad de Parkinson basado en la selección de características del valor SHAP**: en esta aplicación lo que se busca es encontrar los factores más determinantes para el diagnóstico del Parkinson, para lo que se utilizan técnicas de inteligencia artifical explicable como SHAP [10].
- **Explicación de la influencia de los atributos de la enfermedad renal crónica**: en este caso se vuelve a presentar una aplicación de análisis de importancia de características tras emplear un clasificador. Nuevamente, la técnica de importancia de características empleada es SHAP [18].

Capítulo 3

Desarrollo y experimentación

En este capítulo se va a mostrar el proceso realizado para la consecución de los objetivos planteados, esto va a incluir desde una descripción exhaustiva del dataset utilizado, el procesamiento de la base de datos, donde se aplicarán técnicas de reducción de la dimensionalidad, hasta el entrenamiento de modelos, la explicabilidad de los mismos y una fase de experimentación en cuanto a la explicabilidad tras haber aplicado reducción de la dimensionalidad.

3.1. Dataset

El dataset utilizado para realizar este proyecto es el elaborado en el Estudio Toledo de Envejecimiento Saludable (ETES) [3], el cual tiene la finalidad de recopilar datos sobre los modelos de fragilidad existentes, así como del envejecimiento y todos los posibles determinantes del mismo.

El estudio se realizó entre los años 2006 y 2009, con 2488 sujetos de más de 64 años. La recogida de datos se realizó en tres fases:

- **1ª Fase:** se realizó una entrevista de una hora y media de duración, realizada por psicólogos donde se obtuvieron datos de tipo demográfico, de calidad de vida, de comorbilidad, malos hábitos, rasgos depresivos y otros factores determinantes para la salud de los pacientes.
- **2ª Fase:** equipos de enfermeros realizan pruebas a los sujetos, tales como tomas de tensión arterial, frecuencia cardiaca, pruebas de desempeño motor, cognitivas y todas aquellas pruebas que requieren instrumental o personal cualificado.
- **3ª Fase:** se extrajeron muestras de sangre de los sujetos para su posterior análisis y recogida de determinadas medidas.

Tras pedir los datos a las instituciones pertinentes, se concede permiso expreso de manera anónima por parte de los investigadores principales del ETES a los mismos, mediante 3 archivos de tipo csv, cada uno correspondiente a una ola de medidas, y un archivo de tipo docx con la entrevista realizada por los psicólogos, puntuaciones y códigos.

Es importante destacar que en la actualidad ya se han recopilado tres olas de recogidas de datos, dentro de las cuales se realizan las tres fases anteriormente mencionadas. Sin embargo, para este trabajo solo se utilizará la primera, ya que en todas las olas se realizan las mismas mediciones a los mismos pacientes que no hubieran fallecido, y, pese a que se introducen algunos nuevos pacientes, podría haber sesgos por sujetos o por la edad de los mismos.

En los datos correspondientes a la ola 1, se presentan 2488 instancias y un total de 1430 variables, donde se incluyen desde la identificación del sujeto con un código para favorecer el anonimato de los pacientes, hasta cualquier tipo de puntuación obtenida en alguno de los test realizados por los enfermeros.

En este dataset se recogen, aunque con bastantes datos faltantes, todas las medidas recogidas en las fases mencionadas anteriormente, por lo que habrá variables que representan medidas de test, cuestionarios o pruebas diagnósticas, además de medidas de análisis de sangre, como el colesterol o el azúcar y un gran número de fechas, tanto de las pruebas realizadas, como de nacimiento y muerte de pacientes o incluso de comienzo en el programa.

Finalmente, destacar que este estudio está integrado en la Red Temática de Investigación en Envejecimiento y Fragilidad (RETICEF) del Instituto Carlos III, que tiene el aval científico de la Sociedad Española de Medicina Geriátrica (SEMEG) y que, además, surge de la colaboración entre la propia Red de Envejecimiento (RETICEF), la Consejería de Sanidad de Castilla la Mancha y el Servicio de Salud de Castilla la Mancha (SESCAM) [3].

3.2. Creación de modelos predictivos

El objetivo principal de este proyecto es lograr predecir la fragilidad para pacientes con ninguno, uno o dos ítems de la escala FTS-5 y algunas variables que se encuentren en el dataset inicial que sean determinantes, a la vez que se comparan distintos enfoques a la hora de abordar el problema. En los siguientes apartados se va a mostrar el recorrido realizado para la consecución de estos objetivos.

3.2.1. Procesamiento de los datos

En primer lugar, es importante recordar que se ha utilizado el dataset que se explicó en la Sección 3.1, en la que se mencionó que va a estar compuesto por 2488 instancias y 1430 variables. El primer desafío que hay que afrontar es el adecuar los datos para tratar de trabajar con ellos.

La variable objetivo del estudio será la puntuación del test o escala FTS-5 de los sujetos, sin embargo, se presenta la primera dificultad, y es que no todas las instancias cuentan con un valor en esta variable. De hecho, solo 1610 sujetos tienen asignado un valor de la escala, lo que supone un 64,7 % del total y una gran pérdida de datos para entrenar los futuros modelos. Por ello, se va a estudiar si se puede obtener alguna puntuación de la FTS-5 en base a otras variables, ya que bastará con detectar puntuaciones iguales o mayores que 25 para determinar que una persona es frágil [8].

- **Análisis de los ítems de la FTS-5**
  Al ser la escala, la suma de 5 ítems, cuyas puntuaciones también se encuentran en la base de datos inicial, se va a observar cuántas instancias tienen un valor en cada variable de las siguientes: *BMI_FTS5*, *romberg_FTS5*, *pase_FTS5*, *marcha_FTS5*, *grip_FTS5*. Estas variables se corresponden, en orden, a cada uno de los ítems de la escala FTS-5 que se explicaron detalladamente en la sección 2.1.3.2.

Variable	Porcentaje de Datos Faltantes (%)
romberg_FTS5	**32.80**
BMI_FTS5	**22.18**
pase_FTS5	**0.00**
marcha_FTS5	**31.51**
grip_FTS5	**25.36**

Cuadro 3.1: Porcentaje de datos faltantes para cada Variable de la FTS-5.

  En el Cuadro 3.1 se pueden ver los datos faltantes de las variables correspondientes a cada ítem de la escala FTS-5, observando que mientras en la variable *pase_FTS5* no falta ningún dato, en la variable *romberg_FTS5* llegan a faltar casi un tercio de los datos, llegando al 32.80 %.

  Comprobando el número de datos faltantes de cada variable, se puede explicar la gran falta de valores en la variable objetivo, sin embargo, se va a proceder a sumar los valores existentes de estas variables, para ver si en algún caso se llega a 25 o más puntos, lo que probablemente sean personas con una fragilidad bastante clara.

  Una vez hecho, se comprueba que hay 79 instancias que aunque le falten valores en las variables de la FTS-5 y no tengan valor en la variable objetivo, sí que llegan a 25 puntos o más sumando las características que sí tienen valor. De esta forma, se reduce el número de valores faltantes en la variable objetivo hasta tener 1689 instancias que sí tienen valor, lo que supone un 67,88 % del total, un aumento de más de un 5 %.

Una vez se ha procesado la variable objetivo, se transforma cada valor de *FTS5_score* a 0 (No Frágil) o a 1 (Frágil) según corresponda, y se almacenará en otra variable denominada *FRAGIL*, que será la nueva variable objetivo. A continuación, se van a estudiar el resto de variables ya que, muchas de ellas son de complicada aplicación, como puede ser una medida obtenida en un análisis sanguíneo, y que por lo tanto, no es interesante trabajar con ellas para cumplir con los objetivos de este trabajo. Además, también hay muchas fechas que se deben eliminar, variables que simplemente representan la identificación del usuario e incluso variables que representan la misma medida pero divididas en cuartiles o quintiles.

Tras esta primera eliminación de variables, se va a proceder a la transformación de otras para su posterior trabajo.

- **Conversión de columnas a valores concretos**
  Muchas de las variables del dataset pertenecen a respuestas dadas por los sujetos a los cuestionarios. Por ejemplo:

P: ¿Algún médico le dijo que había tenido un accidente cerebrovascular (trombosis, embolia o hemorragia en el cerebro o coágulo)?
R: 1. Sí 2. No 3. NS 4. NC

Al buscar una respuesta únicamente afirmativa para considerar esa variable como positiva, se deben cambiar los valores para que en el dataset si la variable es 2, 3 o 4 se transformen en un 0 (No se confirma accidente cerebrovascular) y el 1 se sigue manteniendo, (Se confirma el accidente cerebrovascular).

Primeramente, se seleccionan las columnas que se quieren transformar y, a continuación, se le pasan los conjuntos de valores originales que se quieren intercambiar por otro valor y una lista con los nuevos valores asignados a cada conjunto. Por último, se le pasa un valor por defecto en caso de que se encontrara un valor que no estuviera definido.
En este caso, el 1 original se intercambiaría por un 1, los valores 2, 3 y 4 por un 0, y en caso de que hubiera otro valor se le asignaría un 0.

- **Creación de variables interesantes**
  Otro paso en el preprocesamiento de los datos, ha sido intentar crear algunas variables a partir de otras para ver si podrían tener un efecto determinante sobre la fragilidad de los sujetos. La idea detrás de esta creación de variables está en los numerosos estudios que se han hecho tratando de relacionar el número de enfermedades crónicas con el estado de fragilidad de los pacientes [12].
  En el dataset inicial no se tiene el número de enfermedades, pero sí cada enfermedad encontrada en el total de sujetos en el estándar CIE-10 [27], en inglés ICD. Por lo que, es fácil conseguir el número de enfermedades que muestra cada paciente. Se puede replicar lo mismo exactamente para los medicamentos, los cuales se encuentran en otro estándar, el ATC [26].

  Quizá, otro ejemplo de variable creada que podría ser interesante para estudiar si tiene relación con el estado de fragilidad actual del paciente, es el consumo de alcohol del sujeto a lo largo de su vida. En el dataset se tienen las siguientes preguntas y respuestas:

  P: ¿Ha bebido alcohol antes?
  R: 1. Sí 2. No 3. NS 4. NC 5. NA

  Si la respuesta es sí, se le realizarán las preguntas del Cuadro 3.2. A partir de aquí, se pueden generar un par de variables, *alcohol_tipo_bebedor* y *alcohol_años_bebiendo*. En el caso de la primera de ellas, se mapean las respuestas en categorías de 0 a 5, siendo 0, no ha bebido nunca, 1 bebedor ligero y 5 bebedor de gran riesgo, y en el caso de la segunda, se corresponde con la resta entre la edad en la que dejó de beber, menos la edad con la que empezó a beber.

  Así se han conseguido dos variables nuevas que indican tanto el consumo de alcohol al día de cada sujeto y durante cuántos años ha estado consumiendo.

  Estos son algunos ejemplos de la creación de variables realizadas para su posterior estudio en relación con el estado de fragilidad actual del paciente.

Pregunta	Opciones	Opciones (cont.)
¿Se podría decir que usted ha sido?	1. Bebedor gran riesgo (H: >13)(M: >8) 2. Bebedor excesivo (H: 9-12)(M: 7-8) 3. Bebedor alto (H: 7-8 UBE/día)(M: 5-6)	4. Bebedor moderado (H: 3-6 UBE/día)(M: 3-4) 5. Bebedor ligero (H: 1-2 UBE/día)(M: 1-2)
¿A qué edad comenzó a beber?	1. <15 años 2. 15-20 años 3. 21-30 años 4. 31-40 años 5. 41-50 años	6. 51-60 años 7. 61-70 años 8. 71-80 años 9. 81-90 años 10. >91 años 11. Sigue bebiendo
¿A qué edad dejó de beber?	1. <15 años 2. 15-20 años 3. 21-30 años 4. 31-40 años 5. 41-50 años	6. 51-60 años 7. 61-70 años 8. 71-80 años 9. 81-90 años 10. >91 años 11. Sigue bebiendo

Cuadro 3.2: Preguntas relacionadas con el consumo de alcohol registradas en el dataset. Se muestra por filas las respuestas posibles a cada pregunta.

Una vez se ha trabajado con las variables, se tiene una primera versión de datos procesados con 1689 instancias y 1142 variables.

El objetivo principal en las siguientes etapas de este proyecto será encontrar cuáles de esas 1142 variables de la primera versión del preprocesamiento son más determinantes a la hora de predecir la fragilidad. Para lograr este objetivo, lo primero que se va a hacer es un análisis exploratorio de los datos (EDA).

3.2.1.1. Análisis Exploratorio de Datos (EDA)

Inicialmente, se va a obtener información sobre el dataset procesado, el cual va a tener un total de 1689 instancias y 1142 características, donde se incluye la variable objetivo *FRAGIL*. Además, todos los valores del dataset procesado van a ser valores numéricos, ya sean números reales o enteros.

A continuación, se debe obtener información sobre los datos faltantes en el dataset. Algunos porcentajes de las primeras variables en orden alfabético y de las últimas son las que se pueden observar en el Cuadro 3.3.

Variable	Valor datos faltantes (%)
ASangretel	28.952
DDlog	36.590
DHEALn	23.091
E_exitoso	13.677
ys5	2.724
ys6	2.901
ys7	2.724
ys8	1.658
ys9	2.072

Cuadro 3.3: Porcentajes de valores faltantes de algunas de las 452 variables a las que al menos les falta un valor. El significado de las variables conocidas puede verse en el Cuadro 1 del Anexo I.

Esto lleva directamente a la conclusión de que hay 452 variables que tienen al menos 1 dato faltante en alguna instancia. Además, hay una gran variedad entre los datos faltantes de unas variables y otras, llegando en ocasiones a faltar casi un tercio de los datos, mientras que en otras ocasiones falta solo un 1 %.

Otro paso fundamental a la hora de realizar el análisis exploratorio de datos, en especial si lo que se quiere es encontrar variables determinantes, es explorar la correlación existente entre las variables predictoras y la variable objetivo. Tras hacerlo, se encuentran los resultados del Cuadro 3.4.

Correlaciones positivas más altas	**Valor**	**Correlaciones negativas más altas**	**Valor**
frailtrailfriedbisi	0.633	cq4	Error
PRE_1	0.630	csinf8a	Error
viña	0.629	fuerza2b	Error
frailtrait	0.628	fuerza3b	Error
frailtrailt	0.628	ppce	Error

Cuadro 3.4: Correlaciones de variables con la variable objetivo. El significado de las variables conocidas puede verse en el Cuadro 1 del Anexo I.

Con estos resultados se pueden extraer dos conclusiones principales, la primera de ellas es, analizando las correlaciones positivas, que si se investiga el significado de las variables, son directamente indicadores de la fragilidad de los sujetos, por lo tanto, no son variables válidas, ya que no es coherente que para predecir la fragilidad se utilice otra medida de fragilidad. La segunda conclusión es, que al obtener correlaciones con valor *Error*, puede deberse a la existencia de variables con un valor constante en todas sus instancias, lo que no aportaría nada de información.

Una vez eliminadas estas variables, se puede observar que ya sí se obtienen valores en las correlaciones negativas:

Correlaciones positivas más altas	Valor	Correlaciones negativas más altas	Valor
frailtrailfriedbisi	0.633	exitosotot	-0.403
PRE_1	0.630	rombc2	-0.409
viña	0.629	fuerza1a	-0.496
frailtrait	0.628	rombc1_correcto	-0.534
frailtrailt	0.628	velocidad_M	-0.540

Cuadro 3.5: Correlaciones de variables con la variable objetivo. El significado de las variables conocidas puede verse en el Cuadro 1 del Anexo I.

Tras solventar el problema de los valores constantes, es hora de abordar el otro problema de variables equivalentes, para ello se va observando la correlación con la variable objetivo, así como las correlaciones de las variables con cada una de las variables de las pruebas de la escala FTS-5 y se va comprobando si realmente son equivalentes o desconocidas, de ser así van siendo eliminadas.

Una vez se han eliminado estas variables, tenemos el primer paso para la creación de los modelos de predicción, y por lo tanto, los datos a utilizar en el entrenamiento de los modelos.

A continuación, se van a calcular las correlaciones de las variables preprocesadas con la variable objetivo, obteniendo así, las 10 variables más correlacionadas con la fragilidad. Una vez preprocesado el dataset inicial, se obtiene:

Correlaciones positivas más altas	Valor	Correlaciones negativas más altas	Valor
marcha_FTS5	0.591	katz1	-0.385
grip_FTS5	0.588	em3	-0.387
romberg_FTS5	0.524	em5	-0.387
ql2	0.453	SILLA_correcto	-0.391
euroql3	0.450	LAWTON6	-0.398

Cuadro 3.6: Correlaciones de variables con la variable objetivo. El significado de las variables conocidas puede verse en el Cuadro 1 del Anexo I.

De estas correlaciones se pueden extraer varias conclusiones principales, destacando que es muy probable que los modelos predictivos que contengan una prueba de la escala FTS-5, la cual sea la velocidad de la marcha, la fuerza de prensión o el test de Romberg obtengan mejores resultados que modelos que contengan solo el IMC o el cuestionario PASE. Además de esto, se van observando otra serie de variables que podrían ser buenas a la hora de predecir el estado de fragilidad al estar muy correlacionadas con esta.

Otra forma de observar el comportamiento de ciertas variables con respecto a la variable objetivo y también demostrar la correlación obtenida, es usando gráficos como el de la Figura 3.1, en la cual se puede observar claramente la correlación altamente negativa entre las variables *LAWTON6* y *FRAGIL*, siendo dominante valores de 1 en *LAWTON6* cuando el sujeto no es frágil y justo al contrario cuando el sujeto es frágil.

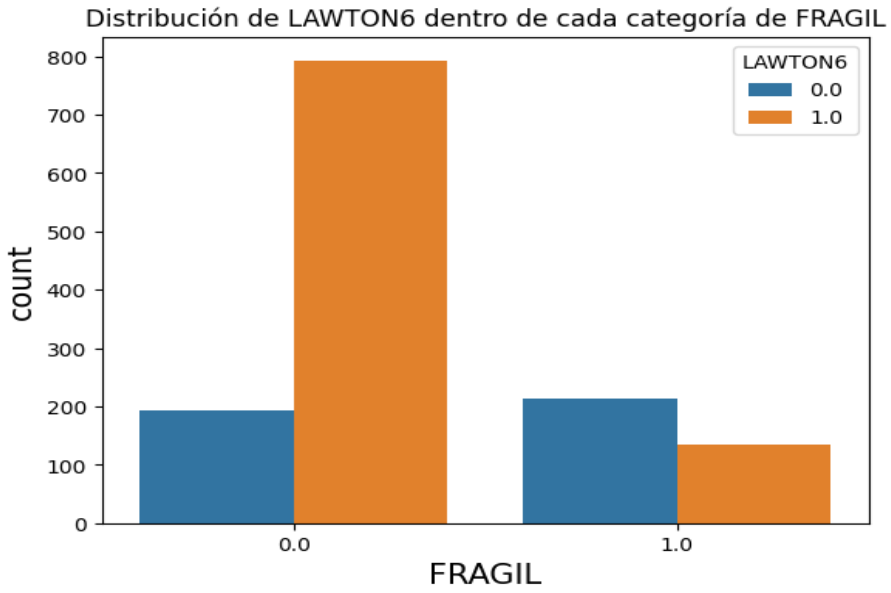

Figura 3.1: Gráfico de distribución de la variable *LAWTON6* en cada categoría de la variable objetivo, *FRAGIL*.

Tras realizar el EDA el número de instancias será 1689 y el número de características para entrenar, 1030.

3.2.1.2. División en conjunto de entrenamiento y test

Respecto a la búsqueda de las características más determinantes para predecir la fragilidad, de la que se hablará en secciones posteriores, es necesario, primeramente, entrenar modelos predictivos.

Antes de proceder a entrenar los modelos, es necesario realizar una serie de adaptaciones a los datos para asegurar el correcto funcionamiento en las predicciones.

En primer lugar, es necesario tener en cuenta que se debe contar con una parte de los datos cuyo objetivo será entrenar el modelo, y otra parte, de menor tamaño, orientada a la evaluación del mismo. Para que la evaluación sea lo más fiable posible, es crucial realizar la separación cuanto antes, evitando así que las manipulaciones que sufran los datos de entrenamiento no afecten a la eficacia de la evaluación. En este caso, se va a realizar una separación de un 80 % de los datos para entrenamiento y un 20 % para datos de test.

Se ha implementado una separación basada en la distribución inicial de la variable objetivo *FRAGIL*, lo que desemboca en una distribución de la variable objetivo en entrenamiento y en el conjunto de test como la que se puede observar en la Figura 3.2. Finalmente, obtenemos un total de 1351 instancias para entrenar el modelo y 338 para evaluarlo, todas ellas con las mismas características.

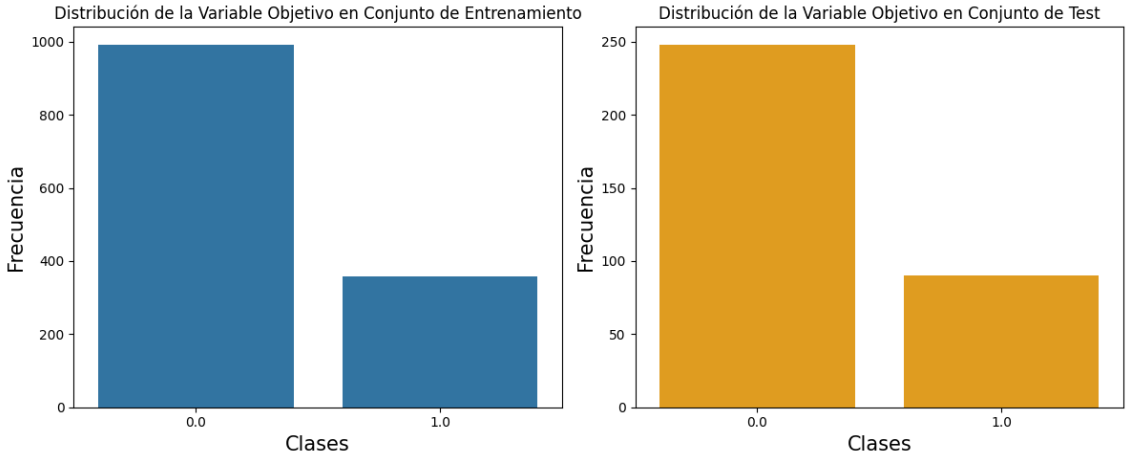

Figura 3.2: Gráfico de distribución de la variable objetivo en los datos de entrenamiento y en los datos de test.

Una vez se ha completado este paso, es hora de contemplar otro gran problema existente en muchos de los dataset utilizados para el entrenamiento de modelos de machine learning, la imputación de datos faltantes.

3.2.1.3. Imputación de datos faltantes

Para tratar los datos faltantes, típicamente se han empleado técnicas como la imputación por media o por mediana, sin embargo, estos métodos son más simples y tienen como principal problema que no son capaces de captar la no linealidad. Por ello, la imputación que se va a emplear en este caso va a ser la imputación por KNN, la cual es más flexible, ya que se pueden modificar ciertos parámetros del método, puede captar relaciones no lineales e incluso puede llegar a ser más eficaz debido a que a la hora de imputar valores faltantes tiene en cuenta las instancias más similares. Al contrario que en la media, por ejemplo, donde una instancia completamente distinta tiene la misma influencia que una muy similar [24].

Es importante destacar que la imputación se realizará por separado al conjunto de entrenamiento y al conjunto de test para evitar que los datos de test influyan en el entrenamiento. Por esto, en el esquema general de la creación de modelos se seguirá por dos caminos diferentes como se muestra en la Figura 3.4, aunque se realicen operaciones similares.

3.2.1.4. Normalización de los datos

Una vez se tiene la certeza que no existe ningún valor faltante en los datos, es necesario escalar los mismos para mejorar el rendimiento de los modelos de machine learning.

La normalización de datos es una técnica que consiste en ajustar la escala de las características de un conjunto de datos. De esta forma, las características presentes van a tener una escala similar, lo que se va a traducir en una mejor eficiencia y eficacia de los algoritmos de machine learning. Esto se debe a que al normalizar los datos, las características van a contribuir por igual al análisis, evitando dar más importancia a características que tengan rangos de valores más amplios. Además, al estar las características en la misma escala, se simplifica la interpretabilidad. En este caso, se va a utilizar la normalización z-score o estandarización. Este tipo de normalización consiste en transformar los datos para que tengan una media de 0 y una desviación estándar de 1 [1].

Esto implica restar la media del conjunto de datos y dividir por la desviación estándar como se muestra en la Fórmula 3.1.

$$z = \frac{x - \mu}{\sigma} \tag{3.1}$$

Donde:

- z es el valor normalizado
- x es el valor original de la característica
- μ es la media de la característica
- σ es la desviación estándar de la característica

En el caso práctico de este proyecto, para poner en funcionamiento esta técnica de normalización se va a utilizar la clase StandarScaler de Python.

Nuevamente es importante destacar cómo esta técnica se aplica de distinta forma al conjunto de entrenamiento y al conjunto de test, por lo que se añade esta fase a ambas ramas presentes en el esquema general como se muestra en la Figura 3.4.

3.2.1.5. Tratamiento de clases no balanceadas

Otro de los problemas que va a aparecer antes de comenzar a entrenar los modelos, es el gran desbalanceo de clases que existe, habiendo un gran número de casos en los que los sujetos no son frágiles, es decir, la variable objetivo es 0, y muy pocos casos en los que los sujetos son frágiles, es decir, la variable *FRAGIL* es 1. Por ello, para que los modelos no tiendan a predecir la clase dominante, en este caso la 0, para obtener la máxima precisión, se va a probar con dos tipos de técnicas.

- **Synthetic Minority Over-sampling Technique (SMOTE)**
  Esta técnica consiste en el aumento de instancias de la clase no dominante, en este caso, sujetos frágiles, pero con la principal característica de que para aumentar este número de instancias, no va a replicar las ya existentes, sino que se van a crear muestras sintéticas. Para crear estas muestras va a utilizar los k vecinos de cada instancia de la clase minoritaria más cercanos y los va a combinar de tal forma que dará lugar a una nueva instancia [2].
- **Random Under-Sampling (RUS)**
  Esta técnica es bastante más simple que la anterior, solo va a consistir en la eliminación de un número de instancias de la clase mayoritaria con el fin de mantener las clases más igualadas.

Además, estas dos técnicas son compatibles y se pueden utilizar secuencialmente. Por norma general, se realizará en primer lugar un RUS seguido de SMOTE, para que la creación de muestras sintéticas no sea tan exagerada y que, a su vez, no se pierda tanta información de la clase mayoritaria al no eliminar tantas instancias.

De esta forma, tras realizar el escalado se podrá elegir entre cuatro opciones distintas, como se muestra en la Figura 3.4, una de ellas sin realizar ningún tipo de muestreo, otra realizando un RUS, realizar un SMOTE, o bien realizar primeramente un RUS y luego un SMOTE. Esto solo se hará en la rama de los datos de entrenamiento, ya que en los datos de test solo se quiere evaluar las instancias ya existentes.

3.2.1.6. Reducción de la dimensionalidad

Como se ha mencionado en apartados anteriores, el número de características a usar es 1030, mientras que el número de instancias es 1689, siendo solo un 80 % de estas para entrenamiento, lo que supone un total de 1351 instancias con 1030 características. Dicho de otra forma, el número de características a la hora de entrenar los modelos es de aproximadamente 3/4 del número de instancias, lo que podría hacer menos eficiente el resultado del mismo al tener una alta dimensionalidad en comparación con el número de muestras.

Por ello, se propone estudiar la aplicación de técnicas de reducción de la dimensionalidad con la intención de mejorar la eficacia de los modelos.

La reducción de la dimensionalidad puede definirse como una técnica de preprocesamiento de datos utilizada para reducir el número de variables con el fin de disminuir el ruido, el coste de computación e incluso aumentar el rendimiento de los algortimos de machine learning.

Alguna de las técnicas de reducción de la dimensionalidad más conocidas son el Principal Components Analysis (PCA) [15] o el Linear Discriminant Analysis (LDA) [4]. Sin embargo, estos métodos tienen algunos inconvenientes, como la falta de localidad, ya que solo tiene una visión global o no tener en cuenta la variable objetivo, en el caso de PCA, o el enfoque únicamente lineal, en ambas técnicas.

Por ello, en este proyecto se va a utilizar otra técnica de reducción de la dimensionalidad, la cual va a solventar estos problemas, la técnica Supervised Local Maximum Variance Preserving (SLMVP) [6].

De esta forma, se aplicará SLMVP como técnica de reducción de la dimensionalidad, tanto en la rama que sigue el procesamiento durante el entrenamiento, como en el camino de test, de tal forma que el esquema general de creación de modelos para predecir la fragilidad quedará como se puede observar en la Figura 3.4, teniendo en cuenta los bloques marcados por líneas discontinuas, siendo justo los que representan la reducción de la dimensionalidad.

3.2.2. Entrenamiento de modelos

Una vez se tienen los datos listos para el entrenamiento, se va a comenzar a buscar el mejor modelo.

Para definir cuál es el mejor modelo, la métrica que se va a utilizar es el f1-score de la clase positiva, por dos razones principales, la primera de ellas es, que al haber clases

desbalanceadas la precisión puede no ser un indicador muy fiable, ya que podría existir un 80 % de precisión en test y no saber detectar un sujeto frágil. La segunda razón es, que al ser un ámbito médico, se va a priorizar el rendimiento de la clase positiva frente a la negativa, ya que es mejor acertar en las condiciones no favorables para el paciente que en las favorables, es decir, se prioriza detectar correctamente si un paciente es frágil realmente, que detectar correctamente cuando no lo es. Será preferible, en cierta medida que existan más falsos positivos que falsos negativos.

A continuación, una vez decidida la métrica a optimizar, se va a utilizar la técnica Stratified K-folds cross-validation (SKCV) durante el entrenamiento.

- **Stratified K-folds cross-validation (SKCV)**
  Esta técnica empleada durante el entrenamiento de los modelos, tiene como objetivo principal encontrar los mejores parámetros para cierto modelo. Su funcionamiento es bastante simple. Primeramente, se divide el conjunto de entrenamiento en subconjuntos, los cuales, en el caso de este tipo concreto de técnica, respetarán la proporción de instancias pertenecientes a cada clase, lo que es bastante óptimo para clases que no están balanceadas.

  En segundo lugar, se irá iterando k veces y seleccionando un subconjunto, también conocido como fold, y se utilizará como datos de validación. El resto de folds se utilizarán para entrenar el modelo. Un esquema de su funcionamiento puede ser el utilizado en la Figura 3.3. De esta forma, se encontrarán los mejores parámetros para un modelo [21].

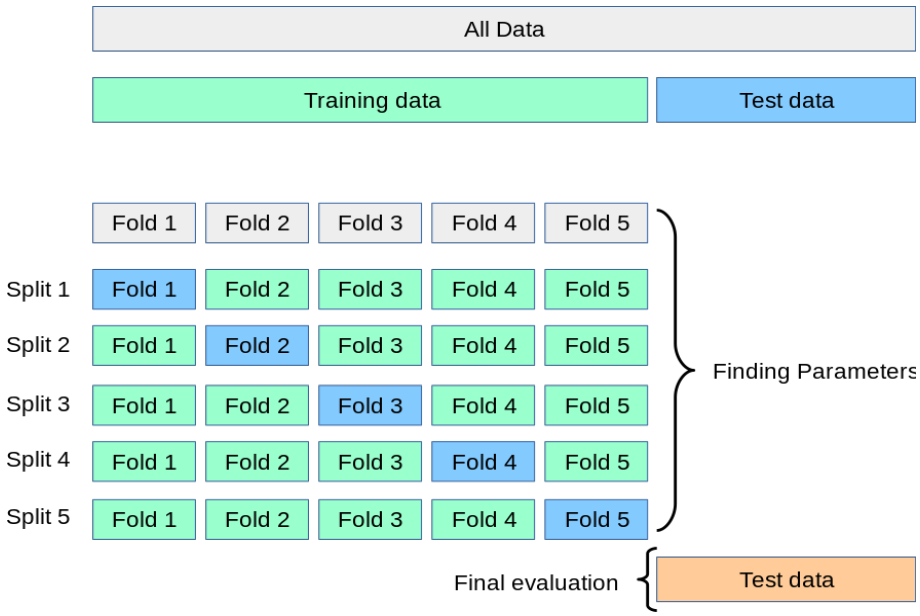

Figura 3.3: Funcionamiento de la técnica SKCV [21].

En el caso de este proyecto, se va a utilizar esta técnica con 5 folds, los cuales buscarán optimizar el f1-score de la clase positiva.

Esto se aplicará a distintos tipos de modelos con la finalidad de ver qué parámetros y de qué tipo de modelo obtiene la mejor f1-score de la clase positiva, considerando este modelo como el mejor. Dicho de otra forma, para cada tipo de modelo se aplicará un SKCV obteniendo los mejores parámetros de cada uno.

Los tipos de modelos a probar se pueden observar en la Figura 3.4 y son los siguientes:

- Support Vector Machine (SVM)
- K-Nearest Neighbors (KNN)
- Extreme Gradient Boosting (XGB)
- Random Forest (RF)
- Ada Boost

Una vez aplicadas estas técnicas, se obtendrán los mejores modelos, es decir, los que tengan mejor puntuación en la f1-score de la clase positiva, y se evaluarán con el conjunto de test, con la finalidad de ver los rendimientos esperados de los modelos para datos no vistos anteriormente. Finalmente, el esquema general de creación de modelos para predicción de la fragilidad quedaría como se muestra en la Figura 3.4, aplicando según interese los bloques de reducción de la dimensionalidad o no aplicándolos para comparar los distintos modelos.

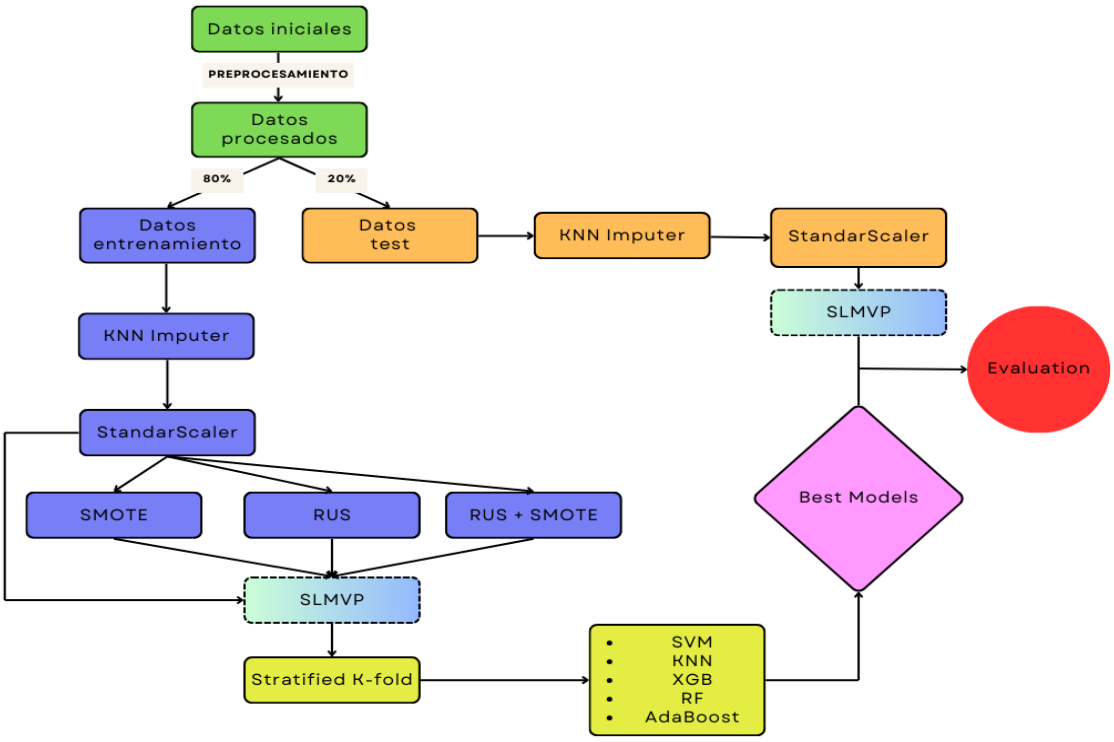

Figura 3.4: Esquema general de creación de modelos para predicción de la fragilidad. Se muestra con línea discontinua la fase de reducción de la dimensionalidad, que solo se aplica en determinados modelos.

A continuación, se van a comparar los resultados obtenidos en cada tipo de modelo, sin aplicar reducción de la dimensionalidad y, los resultados aplicando la misma, en la mejor versión, en cuanto al muestreo de ambas. Esto se hace con la finalidad de encontrar los modelos que mejores predicciones de la fragilidad realicen para posteriormente utilizar técnicas de explicabilidad que puedan aportar algo de transparencia a los modelos y extraer así características relevantes con las que crear modelos de muy baja dimensionalidad.

3.3. Explicabilidad en modelos

Una vez ya se ha hecho el EDA e incluso se han obtenido algunas de las variables que probablemente funcionen mejor a la hora de predecir los modelos finales, se quiere dar otro enfoque a la búsqueda de características relevantes. Para ello, se van a entrenar una serie de modelos con todas las características, y una vez entrenados se extraerán las características más relevantes.

Al entrenar distintos tipos de modelos para comprobar cuál de ellos obtiene mejores resultados, se va a necesitar una técnica de explicabilidad que sea agnóstica de modelos. Uno de los métodos más conocidos y que mejor funcionan para esta tarea es el método SHAP [11].

Para empezar, se comienza con un valor base que se corresponde a la salida promedio de todas las predicciones, en el caso del proyecto, es la probabilidad promedio de que un paciente sea frágil. A continuación, se pone en marcha una permutación de características, es decir, se van eliminando características y se va comprobando cómo varía la predicción.

De esta forma, se va obteniendo la forma en que cada característica contribuye a la predicción, para que finalmente este método agregue todas las contribuciones y se obtenga un valor final, que se corresponderá al valor SHAP de una característica. Finalmente, se podrá saber cuánto de importante es una característica para el modelo y cómo influye en la predicción, en este caso, por ejemplo, si un valor alto en una característica influye en que un sujeto sera más propenso a ser frágil o a que no lo sea [23].

Este método tiene una serie de ventajas como puede ser su flexibilidad, su consistencia o la transparencia que aporta sobre los modelos [11].

Volviendo al caso de este proyecto, cuya finalidad es encontrar variables determinantes para predecir la fragilidad, se va a aplicar esta técnica, entrenando distintos tipos de modelos.

Una vez entrenados, se aplicará la técnica SHAP para extraer la información de la importancia que da cada modelo a cada característica. En el procedimiento que se está siguiendo se van a escoger los dos mejores modelos, ya que son los que, en teoría, predicen con más sentido la fragilidad en los pacientes y por lo tanto, los que asignarán con más sentido la importancia de las características.

Es importante destacar que la elección de los dos mejores modelos es arbitraría, podría haberse escogido solo el mejor o algún modelo más para aplicar técnicas de explicabilidad. En este caso, se han escogido dos con la finalidad de ampliar un poco más el rango de características que los modelos entienden como relevantes para la

posterior creación de modelos con pocas características, y a su vez comprobar si dos modelos distintos con puntuaciones en la evaluación similares asignan importancias parecidas a las mismas características.

3.4. Explicabilidad en modelos. Reducción de la dimensionalidad

Para buscar las variables más determinantes tras haber aplicado anteriormente reducción de la dimensionalidad, se va a aplicar un procedimiento exploratorio secuencial que comenzará con el entrenamiento de modelos que se puede observar en la Figura 3.4. De esta forma, se obtienen los mejores modelos en cuanto a f1-score de la clase positiva, tal y como se hizo en la Sección 3.2.1.6.

Llegados a este punto, el procedimiento será igual que en la sección anterior, es decir, se aplicará SHAP, aunque en lugar de utilizarlo con las características iniciales, se va a emplear con las dimensiones creadas durante la elaboración del modelo predictivo.

De esta forma, se obtienen las importancias que el modelo asigna a cada dimensión, aunque por norma general, al usar SLMVP la dimensión 0 será la que contenga mayor cantidad de información y por lo tanto, será la más determinante para el modelo.

Sin embargo, aquí se plantea el primer problema de esta solución, las dimensiones no son las características originales. Estas variables son características artificiales creadas a partir de las originales. Por ello, lo que se va a plantear es obtener una medida de correlación de cada característica original con respecto a esta dimensión.

Es importante destacar, que esto podría realizarse con cualquier dimensión, sin embargo, es preferible escoger las que mayor valores SHAP tienen debido a que son las que mayor aportación en la predicción hacen. Una vez se tienen los coeficientes de correlación, se va a pasar un umbral como límite, que será arbitrario para identificar qué variables están más asociadas a estas dimensiones.

A continuación, se va a realizar un entrenamiento con un tipo de modelo de clasificación con las variables más correlacionadas con la dimensión elegida, para posteriormente comparar si ha habido pérdida en la evaluación. Si no la ha habido o no es muy importante, se le aplicará SHAP a este nuevo modelo que sí que tendrá las características originales.

El razonamiento detrás de este proceso, es detectar primeramente qué dimensiones influyen con mayor fuerza en la predicción del modelo inicial, al que se le ha aplicado reducción de la dimensionalidad, siempre y cuando se hayan obtenido puntuaciones que verifican que el modelo no es aleatorio. Las dimensiones resultantes serán, muy probablemente, las que hacen que el modelo no sea aleatorio. Una vez se tienen las dimensiones más influyentes, se intenta con algún método, en este caso una medida de correlación, medir qué variables iniciales están más correlacionadas con las dimensiones influyentes. En este punto, obtenemos variables relacionadas con estas dimensiones. A partir de aquí, se entrenarán modelos con cada grupo de variables para ver si los modelos que se generan no son aleatorios. Si no lo son, se vuelve a aplicar SHAP o alguna otra técnica de explicabilidad para ver dentro de las características más correlacionadas con las dimensiones influyentes, cuáles actúan con más fuerza en la predicción de los nuevos modelos, lo que puede aportar una idea

de las variables que inicialmente tienen más impacto en el modelo al que se le aplicó reducción de la dimensionalidad [16].

3.5. Modelos finales de predicción de la fragilidad

Finalmente, se van a crear los modelos para predecir la fragilidad con un número lo más reducido posible de características, las cuales sean de fácil aplicación y con las que se obtengan modelos que mantengan ciertos niveles de precisión diagnóstica.

Aunque, antes de crear estos modelos, se va a dar un paso breve anterior que es descubrir cómo se comportan los ítems del test FTS-5 por separado junto a un número reducido de características relevantes, con la finalidad de saber qué ítems de esta escala pueden tener un mejor desempeño en los modelos finales.

Para lograr esta tarea, se van a entrenar modelos con pocas características, todas ellas relevantes obtenidas a partir de la explicabilidad o el análisis de datos, y además, se les va a añadir una característica correspondiente a una puntuación de la escala FTS-5, es decir, una de las siguientes variables: *BMI_FTS5*, *grip_FTS5*, *marcha_FTS5*, *pase_FTS5* y *romberg_FTS5*.

3.5.1. Modelos predictivos con pocas características

Una vez hecho esto, se crean los modelos finales siguiendo el esquema de la Figura 3.4, sin tener en cuenta la reducción de la dimensionalidad, ya que estos modelos ya tendrán pocas características.

Estos modelos pueden no incluir ningún ítem de la escala FTS-5 para tratar de hacer predicciones de fragilidad solo con una simple entrevista de teléfono o bien incluir alguno de los ítems para hacer predicciones de fragilidad en consultas donde los recursos sean más limitados y no se pueda realizar el test al completo.

Para desarrollar estos modelos, se seleccionará primeramente un conjunto de características, que habiéndolas estudiado anteriormente, sean relevantes a la hora de predecir el estado de fragilidad de un paciente. Por otra parte, al realizar esta selección, se debe verificar que, en caso de que la característica sea el resultado de una prueba diagnóstica, su aplicación no sea muy compleja y que no requiera instrumental avanzado.

Una vez seleccionado el conjunto de características a utilizar, se seguirá el mismo procedimiento visto en secciones anteriores, el cual se puede visualizar en la Figura 3.4, con la única diferencia de que los datos ya se encuentran preprocesados, ya que se hizo anteriormente. Además, es importante destacar que para la selección final de características, se irá probando con variables utilizadas en la escala FTS-5 y otras variables que no pertenecen a la misma, hasta conseguir los mejores resultados.

Finalmente, se realizará la evaluación de los mejores modelos y se valorará por una parte, aquellos modelos que consigan mejores métricas, en especial la f1-score de la clase positiva, y por otra, aquellos modelos que utilicen un menor número de variables. De esta forma, se obtendrán algunos modelos con distintas combinaciones de características para predecir la fragilidad.

3.6. Metodología

A lo largo de este trabajo se han abordado distintas fases para lograr los objetivos planteados en el proyecto.

En primer lugar, se plantearon los problemas presentes actualmente en la literatura relacionados con la dificultad de la aplicación de la escala FTS-5 en centros no especializados. Acto seguido, se planificó cómo podía abordarse este problema y qué soluciones podrían plantearse para pasar a estimar la duración de las tareas que se iban a desarrollar en los meses próximos.

El primer paso en la ejecución del proyecto fue la recopilación de datos del Estudio Toledo de Envejecimiento Saludable (ETES). Estos datos fueron fundamentales para el desarrollo de los modelos predictivos.

Una vez se obtuvieron los datos, fue necesario analizarlos y manipularlos. El preprocesamiento de datos incluyó varias etapas para asegurar la calidad y consistencia del análisis. Primeramente, se estudiaron detenidamente los datos para entender su naturaleza y el comportamiento de los mismos. Acto seguido, se eliminaron variables irrelevantes que no aportaban información significativa e incluso se llegaron a crear variables intermedias. Luego, se procedió a la imputación de datos faltantes utilizando la imputación por KNN, que permite estimar y llenar los valores perdidos de manera eficiente. Además, se normalizaron los datos mediante el uso de StandardScaler para garantizar que todas las variables tuvieran una escala comparable, lo cual es esencial para el rendimiento óptimo de los algoritmos de aprendizaje automático. Finalmente, se abordó el análisis de clases desbalanceadas, un problema frecuente en estudios médicos, mediante técnicas de sobremuestreo y submuestreo para equilibrar las clases y mejorar la capacidad predictiva de los modelos.

Se exploraron y compararon diversas técnicas de reducción de dimensionalidad, como el PCA, LDA o SLMVP. Estas técnicas se contrastaron con los resultados de algoritmos de machine learning aplicados sin reducción de dimensionalidad.

Más tarde, se trató de buscar las variables más relevantes en los modelos creados, y para ello se hizo especial énfasis en técnicas de explicabilidad, como SHAP para identificarlas y poder usarlas posteriormente.

Los modelos predictivos se desarrollaron usando una variedad de algoritmos de aprendizaje automático, incluyendo SVM, KNN, XGBoost, RandomForest y Ada Boost. Cada uno de estos algoritmos se entrenó utilizando un conjunto reducido de ítems, seleccionados mediante análisis de importancia de características.

Durante el entrenamiento de los modelos, se emplearon técnicas de validación para evaluar su rendimiento y evitar el sobreajuste, como la validación cruzada estratificada K-fold, la cual se utilizó para asegurar que los resultados fueran robustos y generalizables. Se evaluaron diversas métricas de rendimiento, como precisión, sensibilidad y puntuación F1, para comparar el rendimiento de los modelos, haciendo especial hincapié en la puntuación f1 o f1-score de la clase positiva.

Por último, se analizaron los resultados obtenidos y se realizaron conclusiones, llegando a realizar una validación con sujetos reales para demostrar si la aplicación de los modelos finales es viable. Además, se plantearon líneas de mejora y posibles líneas de investigación.

Capítulo 4

Resultados

En este capítulo se van a ir recogiendo los resultados obtenidos a lo largo de las etapas del Capítulo 3, así como explicaciones de los mismos e incluso conclusiones a las que se puede llegar en base a estos. Es importante destacar, que todas las métricas de evaluación de modelos que se muestran en este capítulo, consisten en 388 instancias, o sujetos de la base de datos, de las cuales 248 representan la categoría *FRAGIL* 0 (No Frágil) y 90 la categoría *FRAGIL* 1 (Frágil).

4.1. Comparación de resultados con y sin reducción de la dimensionalidad

En este apartado se van a mostrar los mejores modelos obtenidos tras ser entrenados con un gran número de características, tanto si se ha aplicado reducción de la dimensionalidad, como si no se ha hecho, con la finalidad de detectar los modelos que mejor predigan la fragilidad en los pacientes y así poder extraer las características más relevantes posteriormente.

Los mejores modelos obtenidos son:

Model	Sampling Method	Dimensionality Reduction	Accuracy	F-Measure_1
SVM	RUS + SMOTE	No	0.84	**0.63**
	RUS	100 dim	0.86	**0.66**
KNN	RUS + SMOTE	No	0.78	**0.57**
	RUS + SMOTE	50 dim	0.73	**0.58**
XGB	RUS + SMOTE	No	0.85	**0.72**
	SMOTE	10 dim	0.81	**0.65**
RF	RUS	No	0.82	**0.68**
	RUS	10 dim	0.77	**0.62**
Ada Boost	RUS + SMOTE	No	0.85	**0.71**
	RUS + SMOTE	100 dim	0.80	**0.63**

Cuadro 4.1: Comparación de resultados de diferentes modelos con y sin reducción de la dimensionalidad.

Como se puede observar a lo largo de la comparativa, la reducción de la dimensionalidad, se traduce en mejoras en los dos primeros modelos. Esto puede deberse a que los algoritmos son más simples, en comparación con los tres últimos tipos de modelos que son todos de ensamblaje, lo que podría hacer que estén mejor preparados ante una gran cantidad de dimensiones. No obstante, es muy probable que si se aumentara algo más el número de dimensiones, en especial si este llegase a ser mayor que el número de instancias, el rendimiento de los modelos que anteriormente han sufrido una reducción de la dimensionalidad mejorase, mientras que los modelos que no han sufrido esta reducción tenderían a empeorar. Además, como punto a favor de la reducción de la dimensionalidad, pese a que dificulta la explicabilidad de los modelos, mejora bastante el coste computacional al entrenar con un número bastante menor de características.

4.2. Resultados de explicabilidad en mejores modelos

El mejor modelo, que utiliza XGBoost, tiene las siguientes configuraciones: la proporción de características a considerar al construir cada árbol es del 100 %, la tasa de aprendizaje es de 0.1, la profundidad máxima de los árboles es 3, el número total de estimadores es 200, la proporción de instancias utilizadas para entrenar cada árbol es del 70 % y se empleó RUS + SMOTE.

Ha obtenido los siguientes resultados:

Model	Accuracy	**F-Measure_1**	Precision_1	Recall_1
XGB	0.85	**0.72**	0.73	0.70

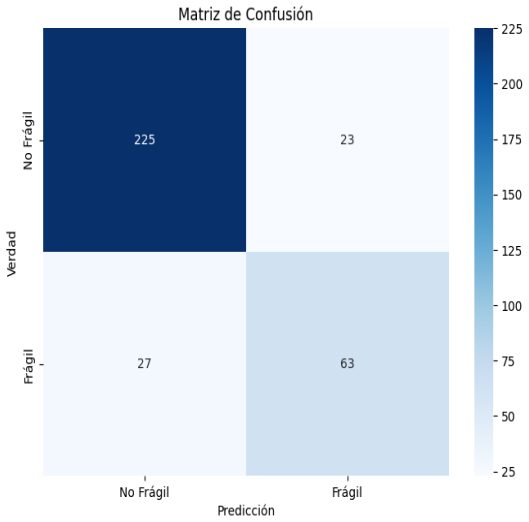

Figura 4.1: Matriz de confusión del modelo XGB.

Tras aplicar SHAP, la importancia de las características puede verse reflejada en la Figura 4.2. Cada punto en la figura representa un paciente, y el color de este punto va a representar el valor de la variable, si tiene un color cálido, el valor será alto,

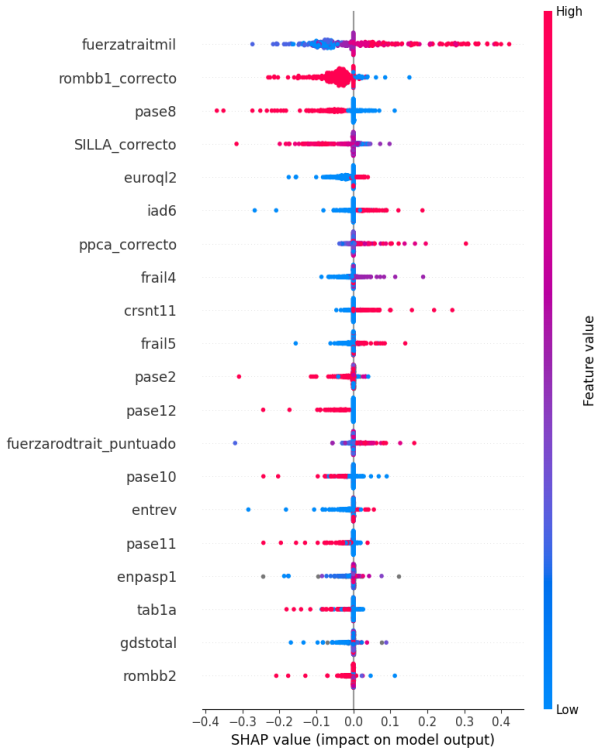

Figura 4.2: Valores SHAP de cada variable en el mejor modelo.

mientras que si tiene un color más frío el valor será más bajo. A medida que se va descendiendo en el orden de características, menor será su impacto en el modelo.

De esta forma, tras obtener el primer modelo se puede deducir que la primera de ellas, *fuerzatraitmil*, es decir, el sesgo de fuerza del paciente, es la que más influye, estableciendo además, que si los valores de la variable son altos, aportará valor para que la predicción sea frágil, mientras que si los valores son bajos, aportará valor para que la predicción tenga como resultado no frágil.

La segunda variable más importante, *euroql2*, correspondiente a la segunda pregunta del cuestionario de calidad de vida europeo, al igual que la tercera, *pase8*, pregunta número ocho del test PASE, funcionan justo al contrario, cuando el valor de la variable es alto, la aportación a la predicción de no frágil aumenta y cuando el valor de la variable es bajo, lo que aumenta es el valor de la predicción frágil.

Llegados a este punto, se tienen las 20 variables más determinantes en la predicción de la fragilidad de los pacientes y cómo estas influyen en la misma.

El segundo mejor modelo, que utiliza Ada Boost, tiene las siguientes configuraciones: la tasa de aprendizaje es de 0.1, la profundidad máxima de los árboles es 3, el número total de estimadores es 100 y se empleó RUS.

Ha obtenido los siguientes resultados:

Model	Accuracy	**F-Measure_1**	Precision_1	Recall_1
Ada Boost	0.83	**0.71**	0.63	0.82

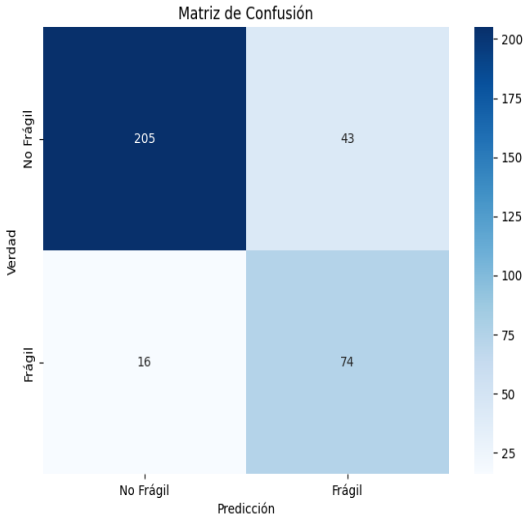

Figura 4.3: Matriz de confusión del modelo Ada Boost.

Los valores SHAP correspondientes a las 20 características más determinantes de este modelo se pueden ver en la Figura 4.4, donde se puede concluir que el modelo funciona de forma diferente al anterior, asignando una influencia mayor a otras características. Aunque, es cierto que muchas características de las 20 más influyentes en ambos modelos coinciden.

De esta forma, se obtiene un abanico más amplio de características con las que poder formar los modelos finales.

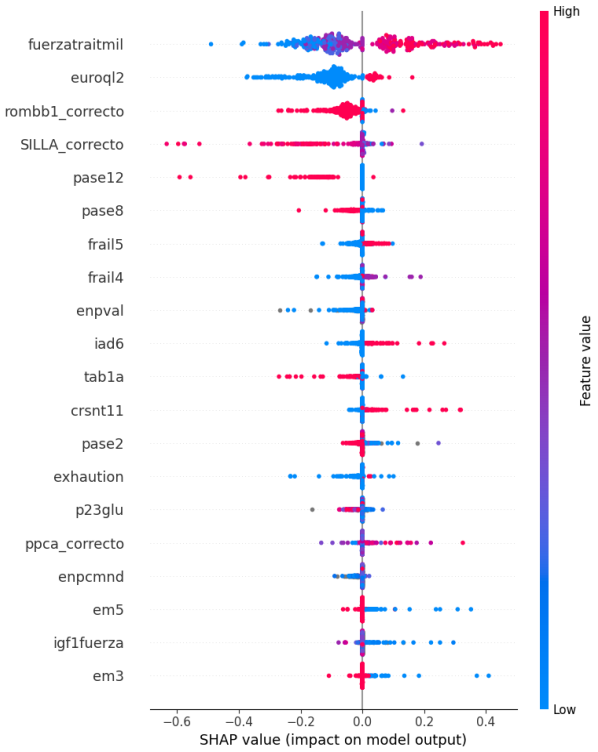

Figura 4.4: Valores SHAP de cada variable en el segundo mejor modelo.

Con la información extraída de la explicabilidad y del análisis de características que se hizo al principio de este trabajo, se pasa a obtener modelos con características seleccionadas, probando con distintas combinaciones de ellas, siempre que se consideren relevantes, cuya aplicación sea posible y, además sean compatibles con otras características, por ejemplo, no se va a combinar la puntuación del test de Romberg con *rombb2* que es una prueba de ese test o no se utilizará la variable *fuerzatraitmil*, pese a ser la más relevante en ambas predicciones, debido a su complicada aplicación.

4.3. Resultados de explicabilidad tras aplicar reducción de la dimensionalidad

En esta sección se van a mostrar los resultados obtenidos al intentar extraer las variables originales más relevantes para un modelo tras haber aplicado reducción de la dimensionalidad.

En este caso, el mejor modelo que se ha obtenido tras haber aplicado reducción de la dimensionalidad es el siguiente:

Model	Accuracy	**F-Measure_1**	Precision_1	Recall_1
SVM (100 dim)	0.86	**0.66**	0.61	0.71

Llegados a este punto, el procedimiento será igual que en la sección anterior, es decir,

se aplicará SHAP, aunque en lugar de utilizarlo con las características iniciales, se va a emplear con las dimensiones creadas durante la elaboración del modelo predictivo.

De esta forma, se obtienen las importancias de la Figura 4.5. Así, se puede conocer qué dimensión ha influido más en la predicción del modelo inicial, siendo la dimensión 2.

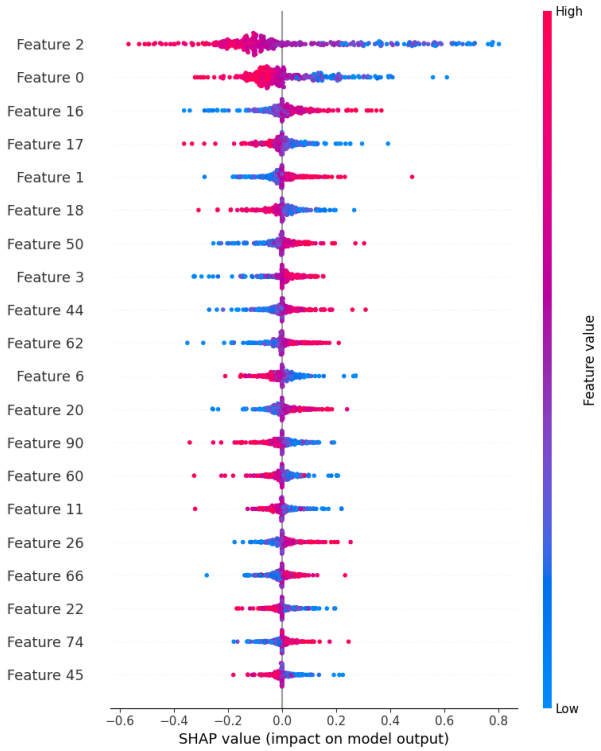

Figura 4.5: Valores SHAP de cada dimensión en el mejor modelo.

En este caso, se puede ver que las dimensiones más influyentes para este modelo son la dimensión 2 y la dimensión 0, pese a que normalmente la dimensión 0 es la que más información contiene. Por esto, a continuación se obtendrán las correlaciones de todas las características del dataset con las dos dimensiones mencionadas.

Tras obtenerlas y aplicar un umbral de 0.3, se observa que hay un total de 75 variables en la dimensión 2 y 32 variables en la dimensión 0 que superan este umbral de correlación con cada dimensión.

A continuación, se va a realizar el entrenamiento con un tipo de modelo de clasificación con las 75 y 32 variables mencionadas en el párrafo anterior y posteriormente, se le aplicará SHAP a cada modelo.

De esta forma, obtenemos los siguientes mejores modelos:

- **Dimensión 2: 75 variables.**

Model	Accuracy	**F-Measure_1**	Precision_1	Recall_1
Ada Boost (SMOTE)	0.78	**0.57**	0.59	0.54

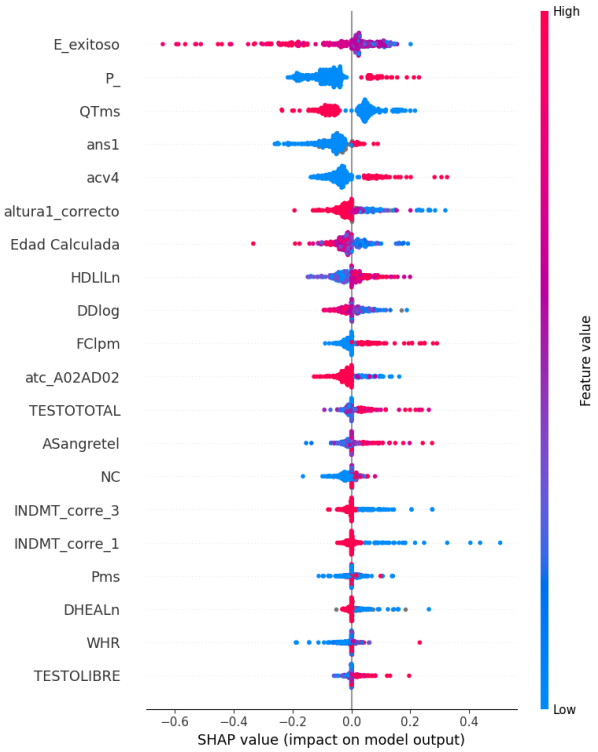

Figura 4.6: Valores SHAP del modelo reducido de la dimensión 0.

En este caso, se ha obtenido un modelo cuyo f1-score para la clase positiva no es demasiado bueno. Es cierto, que el modelo puede ser bueno para predecir la clase negativa y que se podrían obtener importancias para ver qué variables son influyentes en esta predicción. Sin embargo, como a lo largo de este trabajo se le está dando importancia a la clase positiva, no se tendrá en cuenta este modelo.

- **Dimensión 0: 32 variables.**

Model	Accuracy	**F-Measure_1**	Precision_1	Recall_1
Ada Boost (RUS)	0.80	**0.63**	0.60	0.66

Este modelo, al contrario que el anterior, sí parece tener un f1-score en la clase positiva mejor, así que se le aplicará SHAP para ver qué variables han tenido más impacto a la hora de la predicción. Estos valores se pueden observar en la Figura 4.6, viendo así qué variables son más influyentes. De esta forma, se ha pasado de 1030 variables a 32, con una pérdida de solo 0.03 en la f1-score de la clase positiva, e incluso con algo más de procesamiento u obteniendo características de distintas dimensiones podría mejorarse.

Además, si se observa la Figura 4.6 se puede ver cómo son otras variables distintas a las que el modelo da importancia, en comparación con las obtenidas en la sección anterior, y con las que se obtienen buenos resultados, por lo que parecen válidas, lo que aporta aun más características relevantes a la hora de predecir la fragilidad con las que poder crear los modelos finales.

4.4. Comparación de ítems de la escala FTS-5

Para lograr esta tarea, se van a entrenar modelos con pocas características, todas ellas relevantes obtenidas a partir de la explicabilidad o el análisis de datos, y además, se les va a añadir una característica correspondiente a una puntuación de la escala FTS-5, es decir, una de las siguientes variables: *BMI_FTS5*, *grip_FTS5*, *marcha_FTS5*, *pase_FTS5* y *romberg_FTS5*. Las características iniciales a probar que no pertenecen al FTS-5 son *tab1a*, *crsnt11*, *iad6*, *em5*, *euroql2*, *LAWTON6* y *ql2*.

Los resultados obtenidos se muestran en las siguientes tablas:

- **BMI**
  El mejor modelo obtenido ha utilizado Ada Boost y los mejores parámetros son: la tasa de aprendizaje es de 0.1, la profundidad máxima de los árboles es 2, el número total de estimadores es 200 y se empleó RUS.

Model	Accuracy	**F-Measure_1**	Precision_1	Recall_1
SVM	0.79	**0.58**	0.61	0.56
KNN	0.77	**0.55**	0.56	0.54
XGB	0.79	**0.58**	0.62	0.56
RF	0.78	**0.59**	0.59	0.60
ADA	0.78	**0.61**	0.59	0.62

- **Grip**
  El mejor modelo obtenido ha utilizado Random Forest y los mejores parámetros son: cada hoja debe tener al menos 2 muestras, un nodo debe tener al menos 5 muestras para poder ser dividido, la profundidad máxima de los árboles es 10, el número total de estimadores es 300 y se empleó RUS.

Model	Accuracy	**F-Measure_1**	Precision_1	Recall_1
SVM	0.84	**0.69**	0.70	0.69
KNN	0.82	**0.67**	0.65	0.69
XGB	0.83	**0.71**	0.66	0.77
RF	0.84	**0.72**	0.67	0.78
ADA	0.85	**0.71**	0.71	0.72

- **Marcha**
  El mejor modelo obtenido ha utilizado Extreme Gradient Boosting y los mejores parámetros son: la proporción de características a considerar al construir cada árbol es del 70 %, la tasa de aprendizaje es de 0.01, la profundidad máxima de los árboles es 7, el número total de estimadores es 100, la proporción de instancias utilizadas para entrenar cada árbol es del 70 % y se empleó RUS.

Model	Accuracy	**F-Measure_1**	Precision_1	Recall_1
SVM	0.85	**0.76**	0.67	0.88
KNN	0.84	**0.72**	0.68	0.76
XGB	0.86	**0.77**	0.71	0.83
RF	0.84	**0.75**	0.66	0.87
ADA	0.86	**0.75**	0.70	0.80

- **PASE**
  El mejor modelo obtenido ha utilizado Support Vector Machine y los mejores parámetros son: la penalización por errores de clasificación es 1, el alcance de influencia de un ejemplo de entrenamiento es 0.01, se utilizó el kernel Radial Basis Function (RBF) y se empleó RUS.

Model	Accuracy	**F-Measure_1**	Precision_1	Recall_1
SVM	0.79	**0.63**	0.60	0.67
KNN	0.79	**0.62**	0.60	0.63
XGB	0.74	**0.61**	0.51	0.74
RF	0.77	**0.60**	0.55	0.67
ADA	0.78	**0.61**	0.58	0.64

- **Romberg**
  El mejor modelo obtenido ha utilizado Extreme Gradient Boosting y los mejores parámetros son: la proporción de características a considerar al construir cada árbol es del 70 %, la tasa de aprendizaje es de 0.01, la profundidad máxima de los árboles es 3, el número total de estimadores es 200, la proporción de instancias utilizadas para entrenar cada árbol es del 70 % y se empleó RUS + SMOTE.

Model	Accuracy	**F-Measure_1**	Precision_1	Recall_1
SVM	0.82	**0.69**	0.64	0.76
KNN	0.80	**0.66**	0.62	0.71
XGB	0.83	**0.70**	0.67	0.72
RF	0.82	**0.68**	0.66	0.70
ADA	0.82	**0.68**	0.63	0.74

Una vez vistos los resultados de los modelos entrenados con cada ítem del test FTS-5, se puede concluir que, los mejores resultados se dan en la marcha (de un 0.72 de f1-score a un 0.75) y en el grip (de un 0.67 de f1-score a un 0.72). Por otra parte, se tienen los otros tres ítems del test que parecen ser algo menos determinantes, el BMI (0.55 a 0.61), el PASE (0.60 a 0.63) y el Romberg (0.68 a 0.70), aunque este último es más influyente que los anteriores.

Como principal inconveniente de estos resultados, se presenta que el grip es la prueba más difícil de medir, ya que se necesitaría un dispositivo como un dinamómetro capaz de evaluar la fuerza de prensión. Esto deja como principal prueba la velocidad de la marcha, ya que aunque no sea tan precisa de medir sin un dispositivo destinado a ello, puede ser obtenida por un evaluador con bastante fiabilidad con un simple cronómetro.

4.5. Modelos finales

En este apartado se van a mostrar los modelos finales con un número reducido de características que se usarán para la predicción de la fragilidad, para ello se va a hacer una distinción en categorías dependiendo del número de ítems de la escala FTS-5 que se usen.

En esta sección se van a observar nombres de variables cuyo significado se mostrará en el Cuadro 2 del Anexo I. En cada modelo se mostrarán las variables que se emplean para la detección de la fragilidad, siendo estas las consideradas finalmente relevantes en cada caso.

4.5.1. Ningún ítem de la escala FTS-5

En esta subsección, el objetivo es mostrar los mejores modelos conseguidos sin emplear ningún ítem de la escala FTS-5. Estos modelos son los indicados para una detección precoz de la fragilidad mediante preguntas de tipo cuestionario que puedan realizarse incluso telefónicamente. Son ideales para un primer barrido de la población. Se presentan los siguientes modelos:

- Mayor puntuación en métricas

  Variables: 'crsnt11', 'euroql2', 'LAWTON6', 'sueno2', 'iad6', 'Edad Calculada', 'frail4', 'pase2', 'pase8', 'pase12', 'frail5', 'altura1_correcto'.

  El mejor modelo obtenido ha utilizado Support Vector Machine y los mejores parámetros son: la penalización por errores de clasificación es 100, el alcance de influencia de un ejemplo de entrenamiento es 0.001, se utilizó el kernel Radial Basis Function (RBF) y se empleó RUS.

Model	Accuracy	**F-Measure_1**	Precision_1	Recall_1
SVM	0.82	**0.67**	0.64	0.71

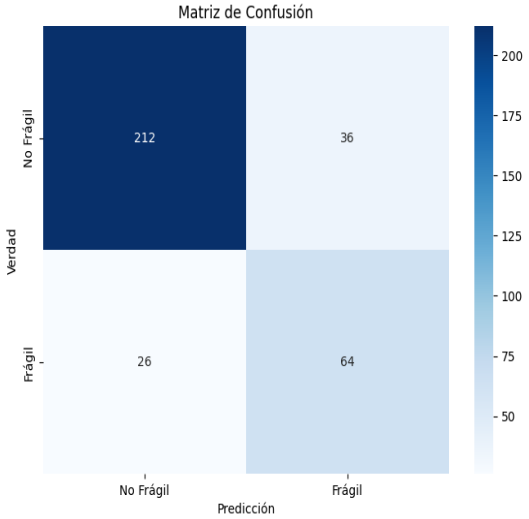

Figura 4.7: Matriz de confusión del modelo SVM.

- Reducción de variables

  Variables: 'frail4', 'iad6', 'Edad Calculada', 'LAWTON6', 'altura1_correcto', 'crsnt11'.

  El mejor modelo obtenido ha utilizado Extreme Gradient Boosting y los mejores parámetros son: la proporción de características a considerar al construir cada árbol es del 70 %, la tasa de aprendizaje es de 0.01, la profundidad máxima de los árboles es 3, el número total de estimadores es 100, la proporción de instancias utilizadas para entrenar cada árbol es del 70 % y se empleó RUS + SMOTE.

Model	Accuracy	**F-Measure_1**	Precision_1	Recall_1
XGB	0.79	**0.64**	0.59	0.70

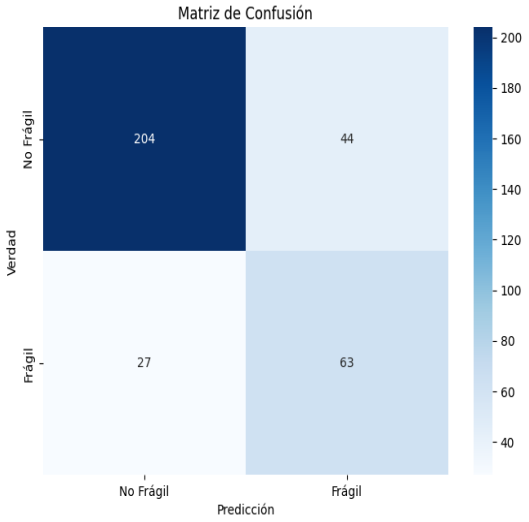

Figura 4.8: Matriz de confusión del modelo Extreme Gradient Boosting.

Como se puede observar en este apartado, el f1-score de ambos modelos no es demasiado alto, aunque sí es suficiente para realizar un primer cribado poblacional, ya que el *recall* de ambos modelos es del 70 % o más. Además, es importante destacar que hay una pérdida significativa en la *precision* al disminuir el número de características a emplear. Sin embargo, al ser su finalidad la evaluación mediante una llamada telefónica, en ocasiones, se priorizará el que esta sea corta a costa de perder puntuación en algunas métricas.

4.5.2. Un ítem de la escala FTS-5

En esta subsección, al contrario que en la anterior, se van a mostrar los mejores modelos conseguidos empleando un ítem de la escala FTS-5, el ítem elegido en este caso será la velocidad de la marcha o *marcha_FTS5*, ya que como se vio anteriormente es la que representa una mejor relación entre eficacia y facilidad de aplicación. Estos modelos están, por lo tanto, pensados para su implantación en consultas de atención primaria, donde no existan recursos para completar el test FTS-5 al completo, o incluso en casas donde los pacientes tengan monitorizado mediante sensores la velocidad de la marcha. Se presentan los siguientes modelos:

- Mayor puntuación en métricas

  Variables: 'marcha_FTS5', 'tab1a', 'crsnt11', 'euroql2', 'LAWTON6', 'ql2', 'euroql3', 'sueno2', 'iad6', 'Edad Calculada', 'dssq15', 'katz1', 'frail4'.

  El mejor modelo obtenido ha utilizado Support Vector Machine y los mejores parámetros son: la penalización por errores de clasificación es 1, el alcance de influencia de un ejemplo de entrenamiento es 0.01, se utilizó el kernel Radial Basis Function (RBF) y se empleó RUS + SMOTE.

Model	Accuracy	**F-Measure_1**	Precision_1	Recall_1
SVM	0.88	**0.79**	0.71	0.89

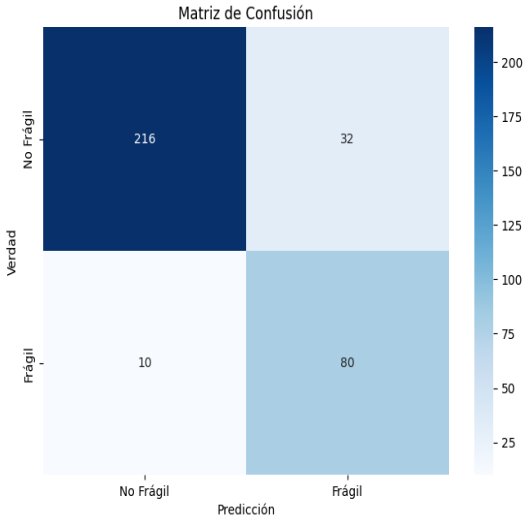

Figura 4.9: Matriz de confusión del modelo SVM.

- Reducción de variables

  Variables: 'marcha_FTS5', 'tab1a', 'crsnt11', 'euroql3', 'sueno2', 'Edad Calculada', 'frail4'.

  El mejor modelo obtenido ha utilizado Ada Boost y los mejores parámetros son: la tasa de aprendizaje es de 1, la profundidad máxima de los árboles es 1, el número total de estimadores es 100 y se empleó RUS + SMOTE.

Model	Accuracy	**F-Measure_1**	Precision_1	Recall_1
Ada Boost	0.87	**0.78**	0.72	0.86

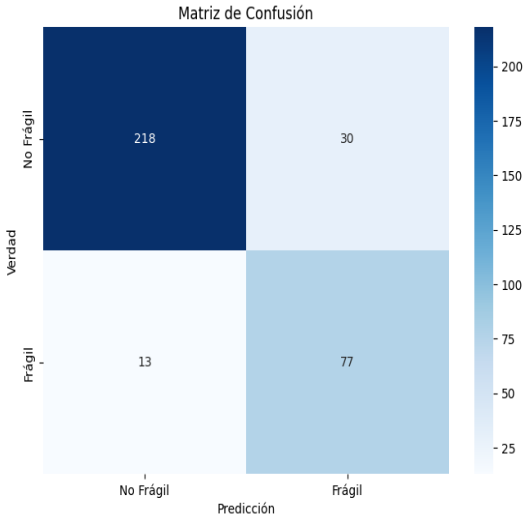

Figura 4.10: Matriz de confusión del modelo Ada Boost.

Como conclusión de esta subsección, se destaca la principal mejora del f1-score, en ambos modelos con respecto a los de la subsección anterior. Esto se traduce, en una solución eficaz para la predicción de la fragilidad al llegar a puntuaciones de f1-score de 0.78 o 0.79. Además, se observa que no hay un gran cambio de métricas al eliminar ciertas variables, por lo que el empleo del modelo con un menor número de variables es preferible.

4.5.3. Más de un ítem de la escala FTS-5

En esta última sección se muestran los resultados obtenidos al tener más de un ítem de la escala FTS-5, estos modelos son menos útiles en cuanto a su aplicación, pero puede darse la situación que se quiera aplicar alguna prueba más de la escala para aumentar las métricas de evaluación y obtener resultados mejores. Para obtener los modelos, se cogen los de la sección anterior y se le añade otro ítem de la escala FTS-5, en este caso será el PASE o *pase_FTS*, ya que es el que mejor resultados da, siendo fácil de aplicar. Este modelo es:

Variables: 'marcha_FTS5', 'pase_FTS5', 'tab1a', 'crsnt11', 'euroql2', 'LAWTON6', 'ql2', 'euroql3', 'sueno2', 'iad6', 'Edad Calculada', 'dssq15', 'katz1', 'frail4',

El mejor modelo obtenido ha utilizado Random Forest y los mejores parámetros son: cada hoja debe tener al menos 2 muestras, un nodo debe tener al menos 10 muestras para poder ser dividido, el número total de estimadores es 100 y se empleó RUS + SMOTE.

Model	Accuracy	**F-Measure_1**	Precision_1	Recall_1
RF	0.89	**0.79**	0.80	0.78

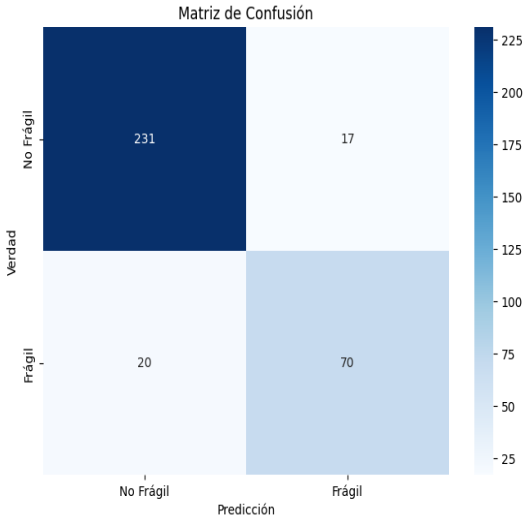

Figura 4.11: Matriz de confusión del modelo Random Forest.

En este caso, se observa como aunque se ha añadido una variable más de la escala FTS-5 la mejora es mínima con respecto a cuando solo se presentaba la velocidad de la marcha, por lo que aplicar dos variables de la escala FTS-5 a los modelos finales no es rentable, en cuanto a que aumenta la complejidad de la evaluación con una mejora mínima en las métricas.

Capítulo 5

Validación

Además del trabajo realizado a lo largo del proyecto, se ha completado una sesión de validación, en el ámbito de un proyecto nacional y aprobado en el comité de ética del Hospital Universitario de Getafe, MOTIVA [14].

Esta sesión se llevo a cabo con un equipo multidisciplinar, formado por una médica de atención primaria, un técnico de salud pública, una enfermera investigadora, un analista de datos y un fisioterapeuta investigador del Hospital Universitario de Getafe en Madrid, a los que se les pasó el consentimiento informado del Anexo II. El objetivo principal fue presentar y discutir los resultados obtenidos, identificar posibles mejoras y explorar los campos de aplicación de las soluciones propuestas.

En esta sesión se presentaron algunos de los resultados obtenidos a lo largo del proyecto para poner en marcha una discusión sobre los mismos, posibles mejoras y campos de aplicación. Se abordaron temas como la precisión diagnóstica, la facilidad de implementación en entornos clínicos con recursos limitados y la posible integración de estas soluciones en la práctica clínica diaria.

Tras la sesión se le pidió a cada asistente que rellenara un cuestionario anónimo y muy breve en relación a las soluciones presentadas. Este cuestionario se diseñó para obtener retroalimentación específica sobre la utilidad, innovación y aplicabilidad de las soluciones, así como sobre cualquier preocupación en cuanto a la seguridad de los pacientes. Los resultados del cuestionario se recopilaron y analizaron para obtener una visión global de las percepciones de los miembros del equipo multidisciplinar.

Los resultados de este cuestionario se pueden observar en el Cuadro 5.1, donde se presenta cada respuesta de cada miembro, siendo 1 completamente en desacuerdo, 5 completamente de acuerdo y NA no aplica, además de la media de las respuestas y la moda de las mismas, para la visión global de los resultados de la validación.

Pregunta	**Usuario 1**	**Usuario 2**	**Usuario 3**	**Usuario 4**	**Usuario 5**	**Media**	**Moda**
Creo que la solución propuesta es útil	4	2	4	4	4	**3.6**	**4**
Creo que la solución propuesta es innovadora	5	5	5	5	3	**4.6**	**5**
Creo que utilizaría la solución en mi práctica habitual	4	1	4	NA	4	**3.25**	**4**
Recomendaría la implementación de la solución en el sistema de salud	3	1	4	4	4	**3.2**	**4**
Creo que la solución propuesta facilitará el trabajo de los profesionales de la salud	5	1	5	4	3	**3.6**	**5**
Creo que la solución propuesta pone en riesgo la seguridad de los pacientes	1	1	1	1	1	**1.0**	**1**

Cuadro 5.1: Resultados de la evaluación de la solución propuesta, siendo 1 Completamente en desacuerdo, 2 En desacuerdo, 3 Neutral, 4 De acuerdo, 5 Completamente de acuerdo y NA no aplica.

En cuanto a la **utilidad de la solución propuesta**, la media de 3.6 y la moda de 4 sugieren que la mayoría de los usuarios creen que las soluciones son útiles, con un usuario manifestando desacuerdo.

La **innovación de la solución** es percibida como altamente innovadora, con una media de 4.6 y una moda de 5, excepto por un usuario al que no le entusiasmaron las soluciones en cuanto a innovación.

En la pregunta sobre **uso en práctica habitual**, la media de 3.25 y la moda de 4 indican una aceptación moderada, aunque un usuario no usaría las soluciones y para otro no aplicaba (NA).

La **recomendación para implementación** tiene una media de 3.2 y una moda de 4, reflejando una opinión positiva sobre recomendar estas soluciones, con un usuario en desacuerdo.

Respecto a si la **solución facilitará el trabajo de los profesionales de salud**, la media de 3.6 y la moda de 5 demuestran que la mayoría cree que las soluciones presentadas facilitarían su trabajo, aunque un usuario no está de acuerdo.

Finalmente, en cuanto a la **seguridad del paciente**, la unanimidad con una media y moda de 1 indica que todos los usuarios coinciden en que las soluciones no ponen en riesgo la seguridad del paciente.

Además, de estas métricas recogidas en el cuestionario, también se les pidió opinión en cuanto a los aspectos positivos de las soluciones, siendo estos: la posibilidad de utilizar la solución en la práctica clínica habitual y la capacidad de detectar variables que, a priori, no considerarían para detectar fragilidad, la alta sensibilidad obtenida, ya que puede ser una herramienta potencial para identificar el riesgo de sufrir efectos adversos como caídas y discapacidades y, además, el uso de técnicas no usadas habitualmente y la innovación de la solución, rápida y fácil de implementar en la consulta de atención primaria.

En cuanto a las posibles mejoras que los usuarios sugirieron para que las soluciones fueran más aceptadas, en especial en términos de usabilidad fueron: la evaluación del algoritmo a nivel incidente y considerar un enfoque incremental que permita descartar fragilidad con las primeras preguntas, que exista un equilibrio entre sensibilidad y facilidad de uso, priorizando la posibilidad de realizar la evaluación de forma no presencial, la reducción del cuestionario para evitar preguntas similares y que estas sean más directas y sin posibilidad de malinterpretaciones o agrupar las preguntas por actividad.

En conclusión, las soluciones propuestas son consideradas útiles e innovadoras, con buena aceptación para su implementación y uso, y son percibidas como seguras para los pacientes.

Estos resultados respaldan la viabilidad de las soluciones en entornos clínicos, aunque hay áreas que podrían requerir más atención y mejora para lograr un consenso total entre los usuarios.

Capítulo 6

Conclusiones

Este proyecto ha abordado la problemática de la fragilidad en personas mayores mediante el desarrollo de modelos de inteligencia artificial, con un enfoque en la explicabilidad y en la integración en la práctica clínica. La fragilidad requiere herramientas diagnósticas precisas y accesibles. A través de este trabajo, se ha buscado mejorar la detección de la fragilidad utilizando modelos de machine learning.

Uno de los aspectos clave del estudio, ha sido la implementación de diferentes modelos como SVM, KNN, XGBoost, Random Forest y Ada Boost, tanto con reducción de la dimensionalidad como sin ella. Los resultados muestran que algunos modelos, alcanzaron una alta precisión diagnóstica y en especial un buen f1-score.

El análisis de explicabilidad realizado mediante SHAP, permitió identificar las variables más influyentes en las predicciones de los modelos. Esto es crucial para la detección de factores influyentes en la fragilidad y en la aceptación clínica, ya que los profesionales de la salud necesitan que los modelos sean lo más transparentes posibles.

Una vez detectados los factores determinantes para la fragilidad, se han creado modelos para la detección precoz de la misma, algunos destinados para la evaluación no presencial, y otros mediante la evaluación en consulta. Los resultados han demostrado que la evaluación sin ningún ítem de la escala FTS-5, es decir, la evaluación telefónica puede tener métricas bastante aceptables, en especial para realizar un primer cribado a la población de riesgo, y que el empleo de la velocidad de la marcha mejora bastante el desempeño de los modelos, mientras que la adición de más variables de la escala FTS-5 deja de mejorar los resultados obtenidos. Dicho de otra forma, se obtienen modelos aceptables para un cribado poblacional y otros modelos a los que se les añade la velocidad de la marcha para su evaluación en consulta.

Además, se realizó una sesión de validación con un equipo multidisciplinar que proporcionó una retroalimentación sobre la utilidad, innovación y aplicabilidad de las soluciones propuestas, además de la evaluación de riesgo de las mismas sobre los pacientes. La mayoría de los participantes consideraron que las soluciones son útiles e innovadoras, y que facilitarían el trabajo de los profesionales de la salud sin poner en riesgo la seguridad del paciente.

Para futuras investigaciones, se sugiere explorar el uso de otras técnicas de inteligencia artificial para mejorar aún más la precisión diagnóstica. Además, sería beneficio-

so investigar la integración de estos modelos en sistemas de salud electrónicos para permitir un monitoreo continuo y personalizado de los pacientes mayores.

En conclusión, este proyecto ha demostrado el potencial de los modelos de inteligencia artificial para mejorar la detección de la fragilidad en personas mayores, ofreciendo herramientas explicables y precisas que pueden integrarse en la práctica clínica. Las mejoras continuas y la validación en entornos reales serán cruciales para su adopción generalizada y su impacto en la salud pública.

Bibliografía

[1] Agarwal, S. (2013). Data Mining: Data Mining Concepts and Techniques. *International Conference on Machine Intelligence and Research Advancement*, 203-207.

[2] Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, 16, 321-357.

[3] ESTUDIO TOLEDO ENVEJECIMIENTO SALUDABLE. (n.d.). ESTUDIO TOLEDO ENVEJECIMIENTO SALUDABLE. http://estudiotoledo.com/

[4] Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7(2), 179-188. https://doi.org/10.1111/j.1469-1809.1936.tb02137.x.

[5] Fried, L. P., Tangen, C. M., Walston, J., Newman, A. B., Hirsch, C., Gottdiener, J., Seeman, T., Tracy, R., Kop, W. J., Burke, G., and McBurnie, M. A. (2001). Frailty in Older Adults: Evidence for a Phenotype. *The Journals of Gerontology: Series A*, 56(3), M146-M157. https://doi.org/10.1093/gerona/56.3.M146

[6] García-Cuesta, E., Aler, R., Pózo-Vázquez, D. D., and Galván, I. M. (2023). A combination of supervised dimensionality reduction and learning methods to forecast solar radiation. *Applied Intelligence*, 53(11), 13053-13066.

[7] García-García, F. J., Carcaillon, L., Fernandez-Tresguerres, J., Alfaro, A., Larrion, J. L., Castillo, C., and Rodriguez-Mañas, L. (2014). A New Operational Definition of Frailty: The Frailty Trait Scale. *Journal of the American Medical Directors Association*, 15(5), 371.

[8] García-García, F. J., Carnicero, J. A., Losa-Reyna, J., Alfaro-Acha, A., Castillo-Gallego, C., Rosado-Artalejo, C., Gutiérrez-Ávila, G., and Rodriguez-Mañas, L. (2020). Frailty Trait Scale-Short Form: A Frailty Instrument for Clinical Practice. *Journal of the American Medical Directors Association*, 21(9), 1260-1266.

[9] Lam de Calvo, O. (2010). Fisiología del Síndrome de Fragilidad en el Adulto Mayor. *Rev Méd Cient*, 20(1), 1249-1256.

[10] Liu, Y., Liu, Z., Luo, X., and Zhao, H. (2022). Diagnosis of Parkinson's disease based on SHAP value feature selection. *Biocybernetics and Biomedical Engineering*, 42(3), 856-869.

[11] Lundberg, S. M., and Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30.

[12] Lv, J., Li, R., Yuan, L., Yi, X., Yi, W., Zi-Wei, Y., and Feng-Mei, H. (2022). Research on the frailty status and adverse outcomes of elderly patients with multimorbidity. *BMC Geriatrics*, 22, 560.

[13] Mahesh, T. R., Geman, O., Margala, M., and Guduri, M. (2023). The stratified K-folds cross-validation and class-balancing methods with high-performance ensemble classifiers for breast cancer classification. *Healthcare Analytics*, 4, 100247.

[14] MOTIVA-Es. (n.d.). AgeingLab. https://ageinglab.ctb.upm.es/motiva-es/

[15] Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 2(11), 559-572. https://doi.org/10.1080/14786440109462720.

[16] Penn State University. (n.d.). Interpretabilidad de Modelos de Machine Learning. https://online.stat.psu.edu/stat505/lesson/11/11.4

[17] Prusty, S., Patnaik, S., and Dash, S. K. (2022). SKCV: Stratified K-fold cross-validation on ML classifiers for predicting cervical cancer. *Frontiers in Nanotechnology*, 4, 972421.

[18] Raihan, M. J., Khan, M. A. M., Kee, S. H., and Nahid, A. A. (2023). Detection of the chronic kidney disease using XGBoost classifier and explaining the influence of the attributes on the model using SHAP. *Scientific Reports*, 13(1), 6263.

[19] Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). Why Should I Trust You? Explaining the Predictions of Any Classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135-1144. https://doi.org/10.1145/2939672.2939778.

[20] Rockwood, K., and Mitnitski, A. (2007). Frailty in Relation to the Accumulation of Deficits. *The Journals of Gerontology: Series A*, 62(7), 722-727. https://doi.org/10.1093/gerona/62.7.722

[21] Scikit-learn. (n.d.). Cross-validation: evaluating estimator performance. https://scikit-learn.org/stable/modules/cross_validation.html

[22] Shapley, L. S. (1953). A value for n-person games. *Contributions to the Theory of Games*, 2(28), 307-317.

[23] Skillsbox. (n.d.). SHAP: una librería de Python para la interpretabilidad de modelos de machine learning. https://medium.com/@Skillsbox/shap-una-libreria-de-python-para-la-interpretabilidad-de-modelos\-de-machine-learning-d111c4bed8a8

[24] Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D., and Altman, R. (2001). Missing value estimation methods for DNA microarrays. *Bioinformatics*, 17(6), 520-525.

[25] World Health Organization. (2022). Ageing and Health. https://www.who.int/news-room/fact-sheets/detail/ageing-and-health

[26] World Health Organization. (n.d.). ATC/DDD Index. https://atcddd.fhi.no/

[27] World Health Organization. (n.d.). ICD-10: International Statistical Classification of Diseases and Related Health Problems. https://www.who.int/publications/m/item/official-who-updates-combined-1996-2019-volume-1

Anexo I

Variable	**Explicación**
cq4	Presencia de la enfermedad Lupus en el paciente, mediante cuestinario
DDlog	Medida del dímero D en sangre del paciente
DHEALn	Índice de DHEA (Dehidroepiandrosterona) del paciente
E_exitoso	Envejecimiento exitoso (tener 10 años por encima de la esperanza de vida)
em3	Tercera pregunta de escala de movilidad
em5	Quinta pregunta de escala de movilidad
frailtrait	Sesgo de fragilidad del paciente
frailtrailt	Sesgo de fragilidad del paciente (mismo que *frailtrait*
fuerza1a	Prueba de fuerza del paciente
fuerza2b	Prueba de fuerza del paciente
fuerza3b	Prueba de fuerza del paciente
katz1	Respuesta de la primera pregunta del test de katz (Valoración actividades de la vida diaria)
rombc1_correcto	Prueba del test de romberg
rombc2	Prueba del test de romberg
velocidad_M	Prueba de velocidad de la marcha
ys5	Pregunta 5 del test de Yesavage (Escala de depresión geriátrica)
ys6	Pregunta 6 del test de Yesavage (Escala de depresión geriátrica)
ys7	Pregunta 7 del test de Yesavage (Escala de depresión geriátrica)
ys8	Pregunta 8 del test de Yesavage (Escala de depresión geriátrica)
ys9	Pregunta 9 del test de Yesavage (Escala de depresión geriátrica)

Cuadro 1: Variables con significado conocido y sus explicaciones.

Variable	**Pregunta**	**Posibles Respuestas**
tab1a	¿Ha fumado diariamente (cada día durante al menos 1 año)?	1. Sí 2. No 3. Dudoso
altura1_correcto	¿Cuánto mide usted?	Respuesta numérica
crsnt11	¿Ha tenido usted los pies o los tobillos hinchados?	1. Sí 2. No 3. NS 4. NC
Edad Calculada	¿Cuántos años tiene usted?	Respuesta numérica
dssq15	¿Ayuda usted al cuidado de los niños de algún familiar, amigo, vecino etc?	1. Sí, diariamente 2. Cada semana 3. Cada mes 4. Menos de una vez al mes 5. No, nunca
em5	¿Es capaz de caminar por la calle cuando hace mal tiempo?	1. Sí 2. No 3. NS 4. NC
euroql2	Cuidado personal	1. No tengo problemas para el cuidado personal 2. Tengo algunos problemas para lavarme o vestirme 3. Soy incapaz de lavarme o vestirme
euroql3	Actividades cotidianas	1. No tengo problemas para realizar mis actividades cotidianas 2. Tengo algunos problemas para realizar mis actividades cotidianas 3. Soy incapaz de realizar mis actividades cotidianas
frail4	Por razones de salud o físicas, ¿Tiene usted alguna dificultad en subir 10 escalones?	1. Si, tiene alguna dificultad 2. No, no tiene dificultad 3. No es capaz de subir 4. NS 5. NC
frail5	Por razones de salud o físicas, ¿Tiene usted alguna dificultad en subir o bajar de un coche/autobús?	1. Si, tiene alguna dificultad 2. No, no tiene dificultad 3. No es capaz de subir o bajar 4. NS 5. NC
iad6	¿Puede permanecer de pie alrededor de 15 minutos?	1. Sin dificultad 2. Con dificultad 3. No puede realizar la función

Variable	**Pregunta**	**Posibles Respuestas**
katz1	A la hora de bañarse	1. No recibe asistencia (entra y sale de la bañera por sí mismo, si la bañera es el medio de limpieza habitual) 2. Recibe asistencia al lavar únicamente una parte del cuerpo (espalda o una pierna) 3. Recibe asistencia al lavar más de una parte del cuerpo (o no se lava) 4. NS 5. NC
LAWTON6	Medios de transporte	1. Viaja solo en transporte público o usa su coche 2. Es capaz de coger un taxi, pero no usa otro medio 3. Viaja en transporte público acompañado por otra persona 4. Solo viaja en taxi o automóvil con ayuda de otros 5. No viaja
pase2	En los últimos siete días, ¿Con qué frecuencia caminó fuera de su casa o de su patio por cualquier razón?	1. Nunca (1 o 2 días) 2. Raras veces (1 o 2 días) 3. Algunas veces (3 o 4 días) 4. A menudo (5 a 7 días)
pase8	En los últimos siete días, ¿Ha hecho algún trabajo doméstico pesado o quehaceres tales como pasar el aspirador, fregar suelos, limpiar ventanas o transportar leña?	1. Sí 2. No 3. NS 4. NC
pase12	En los últimos siete días, ¿Ha cuidado usted de otra persona, tal como un niño, cónyuge dependiente o de otro adulto?	1. Sí 2. No 3. NS 4. NC
ql2	Actividades de la vida diaria en la última semana	1. Ha podido comer, lavarse, ir al retrete y vestirse sin ayuda; utilizar transporte público, conducir su propio coche 2. Ha necesitado ayuda (de otras personas o ayudas técnicas) para realizar las actividades de la vida diaria o el transporte, pero ha podido realizar tareas sencillas 3. No ha podido realizar ni el autocuidado ni tareas sencillas y/o no ha salido de casa o de la residencia

Variable	**Pregunta**	**Posibles Respuestas**
sueno2	En las pasadas 4 semanas, ¿Cuántas veces se ha levantado con la sensación de haber descansado lo suficiente?	1. Nunca 2. Casi nunca 3. Algunas veces 4. Bastante a menudo (con bastante frecuencia) 5. Muy a menudo (con mucha frecuencia)

Cuadro 2: Variables tipo pregunta-respuesta empleadas en los modelos finales.

Anexo II

Consentimiento Informado para la participación en el proyecto *"Ecosistema computacional con apoyo motivacional y evaluación funcional para un programa autónomo de ejercicio para un envejecimiento saludable (MOTIVA)"*

Fundación de Investigación Biomédica del Hospital Universitario de Getafe (FIBHUG),

Universidad Politécnica de Madrid (UPM)

Título del estudio: Co-diseño de sistema de monitorización e intervención para el envejecimiento saludable de las personas mayores.

Investigadores de MOTIVA que intervienen en la actividad:

- Ignacio Peinado Martínez (*ignacio.peinado@salud.madrid.org*), miembro del equipo de trabajo del proyecto MOTIVA (FIBHUG) y facilitador de la sesión
- Olga Laosa Zafra (*olga.laosa@salud.madrid.org*), investigadora principal del proyecto MOTIVA (FIBHUG)
- Elena Villalba Mora (*elena.villalba@ctb.upm.es*), investigadora principal del proyecto MOTIVA (UPM)

I. PROPÓSITO DE ESTE ESTUDIO

Este estudio se lleva a cabo en el ámbito del proyecto *MOTIVA*, cuyo objetivo es diseñar un sistema informático completo que permita hacer un seguimiento detallado del estado funcional de las personas mayores por parte del personal sanitario. Dicho personal, además, tendrá la posibilidad de proponer y monitorizar una intervención multicomponente en la que se abordarán tres aspectos esenciales para promover el envejecimiento saludable: el ejercicio físico, la alimentación y la medicación. El usuario recibirá las instrucciones sobre su plan de ejercicio, nutrición y medicación a través de una aplicación móvil, y podrá recoger información objetiva sobre su capacidad funcional mediante una serie de sensores y cuestionarios. El sistema, además, incorporará mecanismos de motivación para incentivar a la persona mayor a hacer el ejercicio físico que le prescriba el personal sanitario. El profesional sanitario realizará el seguimiento de sus pacientes a través de una plataforma web.

Para maximizar la usabilidad, aceptabilidad y adopción de esta tecnología, se va a realizar un proceso de Diseño Centrado en el Usuario (ISO 9241:11), en el cual se tiene en cuenta a las personas usuarias finales, tanto usuarios mayores como profesionales, a lo largo de todo el proceso.

Como parte de un proyecto de fin de máster realizado en la Universidad Politécnica de Madrid, el alumno Adrián Arana Hernández ha entrenado una serie de modelos de Inteligencia Artificial / Aprendizaje de Máquina (IA/AM) para la predicción de la fragilidad en personas mayores a través de datos recogidos habitualmente en la práctica clínica. El principal objetivo de dichas herramientas es permitir realizar un screening de pacientes que permita la identificación temprana de la fragilidad y la pre-fragilidad a través de datos recogidos habitualmente en la práctica clínica.

Siguiendo este espíritu, en este estudio se pretende analizar la viabilidad y utilidad de dichos modelos. Para ello, se va a llevar cabo un *focus group*, en que participarán profesionales de distintos perfiles. El objetivo de la sesión es que los y las participantes debatan y evalúen las fortalezas y las debilidades de los diferentes modelos y soluciones propuestas, basándose en su experiencia y expectativas.

Por todo ello, en la sesión participarán una serie de profesionales de diferentes disciplinas, incluido usted:

- Ignacio Peinado (*Ignacio.peinado@salud.madrid.org*), miembro del equipo de trabajo del proyecto MOTIVA y facilitador de la sesión.
- Ángel Rodríguez Laso (*arodriguezlaso@salud.madrid.org*), técnico en salud pública
- Olga Laosa Zafra (*Olga.laosa@salud.madrid.org*), farmacóloga clínica y médica de primaria.
- José Antonio Carnicero (*lobouc3m@hotmail.com*), analista de datos

En ninguno de los casos se pretende evaluar o juzgar a ninguna de las personas que participen en el estudio, sino debatir y hallar las mejores soluciones para las personas mayores, que serán las usuarias del sistema.

II.PROCEDIMIENTOS

La sesión tendrá lugar de manera presencial el próximo miércoles 26 de junio de 2024 en la sala de reuniones de la Fundación de Investigación Biomédica del Hospital Universitario de Getafe, entre las 11:30 y las 13:30.

Antes de que inicie la sesión, el facilitador se asegurará de que el participante ha leído y comprendido este consentimiento informado y que, tras resolverle todas las dudas que pueda tener, lo firme de manera totalmente consciente y voluntaria.

Tras esto, el facilitador se encargará de explicar al resto de los participantes el objetivo de la sesión y cómo ésta se llevará a cabo.

Durante la sesión, el proyectando realizará una breve presentación de cada uno de los modelos y pedirá a los y las participantes que debatan las fortalezas y las debilidades de cada uno de los modelos, incluyendo la importancia de las diferentes features utilizadas para entrenar a los modelos o la evaluación del rendimiento de cada modelo. Se dispondrán también de hojas de papel en blanco y material de escritura para poder “dibujar” en caso de ser necesario.

Finalmente, el facilitador planteará una serie de preguntas abiertas con el fin de identificar potenciales problemas, limitaciones o aspectos clave a tener en cuenta para asegurar el éxito de la solución en términos de usabilidad, aceptabilidad y adopción de la tecnología.

III. RIESGOS

No se prevé que existan riesgos por participar en esta sesión de co-diseño. En cualquier caso, el equipo de investigación de este estudio tomará todas las medidas que sean necesarias para asegurar que las personas participantes en el estudio, y especialmente las personas mayores, no sufran ningún daño o inconveniente.

IV. BENEFICIOS DE ESTE PROYECTO

Su participación en este estudio ayudará a diseñar una solución que sea útil para los profesionales en su práctica clínica diaria. Esto permitirá que podamos desarrollar un sistema que permita identificar de forma temprana a los pacientes que tienen mayor riesgo de caer en una situación de pre-fragilidad o fragilidad sin necesidad de realizar costosas pruebas presenciales. Además, dichos algoritmos permitirán que el personal sanitario pueda ofrecer un plan de intervención ajustado al estado funcional actual y previsto, actualizado, dotado además de mecanismos de motivación. En última instancia, esto permitirá que las personas mayores robustas mantengan su buen estado de saludad, que aquellas en estado frágil o pre-frágil puedan mejorar su estado de salud y, sobre todo, que todas ellas eviten transitar a un estado de discapacidad.

V. ALCANCE DEL ANONIMATO Y CONFIDENCIALIDAD

Con el objetivo de asegurar la completa recogida de toda la información que se genere durante la sesión, el sonido de ésta será grabado. El facilitador será el encargado de realizar dicha grabación.

El único fin de esta grabación es permitir que las personas que participen en la sesión se concentren en el debate y el co-diseño, y no en la recogida de información. Las grabaciones serán usadas únicamente para realizar una transcripción completa de lo que se diga en la sesión. El acceso a estas grabaciones estará limitado de forma estricta. En primer lugar, tendrá acceso a ellas el facilitador de la sesión, D. Jaime Ramírez Rodríguez, que es quien llevará a cabo la grabación. Una vez finalizada la sesión, la grabación será depositada en el servicio oficial de almacenamiento en la nube de la Universidad Politécnica de Madrid (Microsoft OneDrive), cuyos servidores se encuentran ubicados en Europa, al cual únicamente tendrán acceso mediante contraseña Dña. Elena Villalba Mora, investigadora responsable de este estudio, y D. Ignacio Peinado Martínez, que será el encargado de realizar la transcripción completa del audio, asegurando que los datos sean anónimos. Todas ellas procederán a eliminar la grabación tras depositarla en la nube.

En lo que respecta a las transcripciones de las grabaciones, únicamente las personas indicadas en este consentimiento informado tendrán acceso a ellas. Dicha información, junto con sus datos personales, será siempre totalmente anónima y privada y se asegurará su confidencialidad. Para ello, las transcripciones también serán almacenadas en el servicio oficial de almacenamiento en la nube de la Universidad Politécnica de Madrid (Microsoft OneDrive), cuyo acceso quedará limitado por contraseña únicamente a las personas indicadas en este consentimiento informado.

En última instancia, el análisis de esta información se hará de forma agregada y anónima, siendo imposible identificar quién es el origen de dicha información. Dichos análisis podrán usarse para difundir en cualquier medio (revistas, congresos, eventos de divulgación, entre otros) los resultados del proyecto MOTIVA.

La responsable en última instancia del estudio, y por tanto de la protección de los

datos derivados de éste, será Dña. Olga Laosa Zafra (FIBHUG), investigadora principal del proyecto MOTIVA en el que se enmarca este estudio.

A continuación, se proporciona la información básica sobre protección de datos:

- Responsable de los datos: Elena Villalba Mora
- Finalidad: Co-diseñar soluciones de monitorización e intervención del estado funcional de personas mayores en el contexto del proyecto MOTIVA.
- Derechos: Acceso, rectificación, y supresión de los datos, así como otros derechos, en los términos y con las limitaciones que se indican en la legislación vigente:
  - Reglamento General de Protección de Datos 95/46/CE
  - Ley orgánica 03/2018, de 5 de diciembre, de Protección de Datos Personales y garantía de los derechos digitales

Si en un futuro se quisiese difundir los resultados de este estudio más allá de lo indicado en este Consentimiento Informado, será necesario obtener de nuevo su autorización expresa.

VI. COMPENSACIÓN

Su participación es voluntaria y no remunerada.

VII. LIBERTAD PARA ABANDONAR EL ESTUDIO

Usted es libre de abandonar este estudio en cualquier momento y por cualquier motivo, y sin necesidad de tener que ofrecer ningún motivo o explicación. Dicha retirada no tendrá ningún tipo de consecuencia para usted.

IX. RESPONSABILIDADES Y PERMISO DEL SUJETO

Sí	**No**	**Declaración**
		Confirmo haber leído este consentimiento informado y las condiciones de este proyecto
		Confirmo haber entendido este consentimiento informado y las condiciones de este proyecto
		Confirmo que puedo participar en este estudio
		Me comprometo a cumplir las normas de este proyecto
		Confirmo aceptar voluntariamente y sin ningún tipo de coacción participar en este estudio, teniendo en cuenta lo expuesto en este consentimiento informado
		Acepto que se grabe mi voz durante la sesión, y que se haga uso de ella en los términos establecidos en este consentimiento informado
		Autorizo a que, en un futuro, se pongan en contacto conmigo para participar en otros estudios o actividades relacionadas con el tema de este estudio

Y para que así conste, firmamos a continuación este consentimiento informado:

Participante	**Investigador**
Firma: Fecha: «Fecha de la persona participante en el estudio » Nombre: «Nombre de la persona participante en el estudio » DNI/NIE: «DNI o NIE de la persona participante en el estudio » Email: «Email de la persona participante en el estudio »	Firma: Fecha: 2 Nombre: Elena Villalba Mora DNI/NIE: XXXX Email: *elena.villalba@ctb.upm.es*

Si tiene alguna pregunta sobre este estudio, puede ponerse en contacto con Dña. Olga Laosa Zafra (*olga.laosa@salud.madrid.org*), investigadora responsable de este estudio.