UNIVERSIDAD EUROPEA DE MADRID

ESCUELA DE ARQUITECTURA, INGENIERÍA Y DISEÑO

MÁSTER EN BIG DATA

TRABAJO FIN DE MÁSTER

MODELADO DE RIESGO PERCIBIDO DE DESARROLLAR CÁNCER SEGÚN HÁBITOS AUTODECLARADOS

BRAD IÑAKI PUGLLA JARA

Dirigido por

DR. ÁLVARO MANUEL RODRÍGUEZ RODRÍGUEZ

CURSO 2025-2026

TÍTULO: MODELADO DE RIESGO PERCIBIDO DE DESARROLLAR CÁNCER SEGÚN HÁBITOS AUTODECLARADOS

AUTOR: BRAD PUGLLA JARA

TITULACIÓN: MÁSTER EN BIG DATA

DIRECTOR/ES DEL PROYECTO: Dr. ÁLVARO MANUEL RODRÍGUEZ RODRÍGUEZ

FECHA: ABRIL DE 2026

RESUMEN

Este trabajo desarrolla y evalúa un modelo de clasificación supervisada para predecir la salud percibida negativa en la población adulta española a partir de hábitos autodeclarados y determinantes sociodemográficos, utilizando los microdatos de la Encuesta de Salud de España 2023 (ESdE 2023), elaborada por el Instituto Nacional de Estadística y el Ministerio de Sanidad. Se integraron los cuestionarios de adultos y hogares excluyendo explícitamente las variables de morbilidad clínica para garantizar la validez conceptual del modelo como instrumento basado exclusivamente en determinantes conductuales y sociales. Además, dado que la fuente no incluye una pregunta directa sobre percepción de riesgo oncológico, la salud autopercibida negativa se emplea como indicador indirecto de dicho riesgo, respaldada por evidencia prospectiva que documenta su asociación con mayor incidencia posterior de diagnóstico de cáncer.

Se compararon tres algoritmos de clasificación: regresión logística, Random Forest y XGBoost. Se realizó un entrenamiento mediante validación cruzada estratificada de cinco pliegues y evaluación sobre un conjunto de prueba independiente. La interpretación de variables reveló que la edad es el predictor más potente, seguida de la movilidad cotidiana, el consumo de alcohol, el tipo de actividad física laboral y el sedentarismo.

Los resultados demuestran que es posible predecir la salud percibida negativa a partir exclusivamente de hábitos y determinantes sociales, sin ningún dato clínico, lo que pone de manifiesto la elevada capacidad informativa de los comportamientos autodeclarados sobre el riesgo percibido de enfermedad en la población española adulta.

Palabras clave: salud autopercibida, regresión logística, XGBoost, Random Forest, encuesta de salud, hábitos autodeclarados.

ABSTRACT

This study develops and evaluates a supervised classification model to predict negative self-rated health in the Spanish adult population based on self-reported lifestyle habits and sociodemographic determinants, using microdata from the 2023 Spanish Health Survey (ESdE 2023), produced by the National Statistics Institute and the Ministry of Health. The adult and household questionnaires were integrated, explicitly excluding clinical morbidity variables to ensure the conceptual validity of the model as an instrument based exclusively on behavioral and social determinants. Furthermore, the source does not include a direct question on cancer risk perception, that's why negative self-perceived health is used as an indirect indicator of said risk, supported by prospective evidence that documents its association with a higher subsequent incidence of cancer diagnosis.

Three classification algorithms were compared: logistic regression, Random Forest and XGBoost. Training was performed using stratified five-fold cross-validation and evaluation on an independent test set. SHAP value analysis revealed that age is the most powerful predictor, followed by daily walking for commuting, alcohol consumption, type of occupational physical activity and sedentary behavior.

The results demonstrate that negative self-rated health can be predicted using exclusively self-reported habits and social determinants, without any clinical data, highlighting the high informative capacity of self-declared behaviors for perceived disease risk in the Spanish adult population.

Keywords: self-rated health, logistic regression, XGBoost, Random Forest, health survey, self-reported habits.

AGRADECIMIENTOS

A mi tutor, el Dr. Álvaro Manuel Rodríguez, por su gran paciencia y ayuda en la realización de este proyecto. Su guía fue clave para presentar el mismo.

A mis amigos, que su compañía fue y será motivo de alegrías. No podría pedir más de ellos.

A mi familia, por darme la fuerza para seguir adelante, gracias por no dejar que la presión de una nueva vida me abrume.

Finalmente, agradezco a todos quienes formaron parte directa o indirectamente en la realización de este trabajo.

Dedicatoria

A mi abuelo, Hugo.

TABLA RESUMEN

	DATOS
Nombre y apellidos:	Brad Puglla Jara
Título del proyecto:	MODELADO DE RIESGO PERCIBIDO DE DESARROLLAR CÁNCER SEGÚN HÁBITOS AUTODECLARADOS
Directores del proyecto:	Dr. Álvaro Manuel Rodríguez Rodríguez
El proyecto se ha realizado en colaboración de una empresa o a petición de una empresa:	NO
El proyecto ha implementado un producto: (esta entrada se puede marcar junto a la siguiente)	SI
El proyecto ha consistido en el desarrollo de una investigación o innovación: (esta entrada se puede marcar junto a la anterior)	SI
Objetivo general del proyecto:	Desarrollar y evaluar un modelo de clasificación supervisada capaz de predecir la salud percibida negativa

Índice

Índice de Figuras

Índice de Tablas

Capítulo 1. RESUMEN DEL PROYECTO

1.1 Contexto y justificación

El cáncer constituye uno de los principales problemas de salud pública a nivel mundial. En 2024, por primera vez, el cáncer superó a las enfermedades cardiovasculares como primera causa de muerte en España, con 115.578 fallecimientos (INE, 2025; SEOM-REDECAN, 2026). La salud autopercibida (SAP) constituye en este contexto un indicador de especial relevancia, en especial la percibida negativa, pues se utiliza como indicador indirecto del riesgo percibido de desarrollar cáncer.

El presente trabajo se enmarca en la intersección entre la epidemiología del cáncer, la salud pública y el aprendizaje automático.

1.2 Planteamiento del problema

La pregunta que se aborda en este trabajo es: ¿es posible predecir la salud percibida negativa, como indicador prospectivo del riesgo de desarrollar cáncer, a partir exclusivamente de hábitos autodeclarados y determinantes sociales, sin recurrir a información clínica? La hipótesis de trabajo es que un modelo de clasificación entrenado sobre hábitos autodeclarados puede identificar a los individuos con mayor riesgo percibido de enfermedad oncológica.

1.3 Objetivos del proyecto

El objetivo del presente trabajo consiste en desarrollar y evaluar un modelo de clasificación supervisada capaz de predecir la salud percibida negativa, como aproximación del riesgo percibido de desarrollar cáncer en la población adulta española, a partir de hábitos autodeclarados y variables sociodemográficas registradas en la Encuesta de Salud de España 2023 (INE & Ministerio de Sanidad, 2025).

1.4 Resultados obtenidos

Se entrenaron y evaluaron tres modelos de clasificación, los cuales mostraron una brecha mínima entre la validación cruzada y la evaluación en test, confirmando la ausencia de sobreajuste. El análisis de variables reveló importancia de edad, movilidad cotidiana, el consumo de alcohol, el tipo de actividad física laboral, el sedentarismo y el índice de masa corporal.

1.5 Estructura de la memoria

La memoria se organiza en nueve capítulos. El Capítulo 2 revisa el estado del arte, determinantes de salud modificables asociados al cáncer y aplicaciones de aprendizaje automático asociado. El Capítulo 3 detalla los objetivos del proyecto. En el capítulo 4 se describe el desarrollo completo del trabajo. En el quinto presenta los resultados en relación con la literatura y las limitaciones del estudio, mientras los 6 y 7 recogen las conclusiones y las futuras líneas de trabajo, respectivamente. Los Capítulos 8 y 9 incluyen las referencias bibliográficas y los anexos.

Capítulo 2. ANTECEDENTES / ESTADO DEL ARTE

2.1 Estado del arte

2.1.1 La salud autopercibida como indicador en epidemiología

La SAP es uno de los indicadores de salud más estudiados y validados en la literatura epidemiológica internacional. Su uso se ha extendido debido a su simplicidad de medición y su demostrada validez predictiva. La SAP global es un predictor independiente de mortalidad en prácticamente todos los estudios revisados, incluso tras controlar por numerosos indicadores de salud específicos y otras covariables relevantes (Idler & Benyamini, 1997). Esta propiedad ha sido confirmada en múltiples contextos culturales y grupos de edad, consolidando la SAP como medida estándar en encuestas de salud internacionales.

En el ámbito europeo, la SAP forma parte del Módulo Mínimo Europeo de Salud (MEHM), que la incluye como primera pregunta del módulo de estado de salud general en todas las encuestas armonizadas de la Unión Europea (Eurostat, 2018). El criterio de dicotomización entre valoración positiva (muy bueno + bueno) y negativa (regular + malo + muy malo) está establecido por Eurostat y es el utilizado de forma consistente en la literatura comparativa europea. El estado de salud percibido refleja la percepción que los individuos tienen sobre su propia salud, tanto desde el punto de vista físico como psicológico o sociocultural, y es un buen predictor de la esperanza de vida, de la mortalidad, del desarrollo de enfermedades crónicas y de la utilización de servicios sanitarios (Ministerio de Sanidad, 2017).

La relación entre la SAP y el riesgo oncológico ha sido objeto de estudio en cohortes prospectivas. Estudios longitudinales han documentado que individuos con SAP deteriorada presentan mayor incidencia posterior de diagnóstico de cáncer, incluso tras controlar por factores de riesgo conocidos, lo que sugiere que la SAP captura señales subclínicas de riesgo antes de que sean detectables por medios convencionales (Benyamini & Idler, 1999). Este vínculo prospectivo justifica el uso de la SAP como proxy del riesgo percibido de enfermedad oncológica en estudios basados en encuestas poblacionales.

2.1.2 Determinantes de la salud percibida: hábitos y factores sociales

La literatura sobre determinantes de la SAP es extensa y consistente en identificar un conjunto de factores modificables como predictores robustos. El marco de determinantes sociales de la salud de Dahlgren y Whitehead (1991) organiza estos factores en capas concéntricas que van desde los estilos de vida individuales hasta los condicionantes socioeconómicos y ambientales más amplios, proporcionando el fundamento teórico para la selección de variables en estudios de este tipo.

Entre los determinantes conductuales, la actividad física ocupa un lugar central. El ejercicio físico puede reducir hasta un 30% el riesgo de cáncer de mama, colon, vejiga urinaria, endometrio, esófago y estómago, y casi el 20% el riesgo de mortalidad específica por cáncer (SEOM, 2024), y su asociación con una mejor SAP está ampliamente documentada. El tabaco y el alcohol constituyen los otros dos hábitos con mayor evidencia acumulada: el tabaquismo es el principal

factor modificable asociado al cáncer, presente en el 15,1% de los diagnósticos a escala global, seguido de las infecciones y el consumo de alcohol (Fink et al., 2026). El índice de masa corporal, como indicador de obesidad, y el sedentarismo completan el conjunto de hábitos con mayor impacto documentado tanto sobre la SAP como sobre el riesgo oncológico.

Los determinantes sociales muestran gradientes consistentes con la SAP en todas las encuestas europeas. La educación, por su capacidad para identificar distintos estratos sociales, es un factor relevante en el proceso de autovaloración de la salud, y los patrones de determinantes difieren significativamente entre grupos educativos y entre hombres y mujeres (Gumà et al., 2023).

2.1.3 Aprendizaje automático en predicción de salud

La aplicación de técnicas de aprendizaje automático en la predicción de resultados de salud a partir de datos de encuestas poblacionales ha experimentado un crecimiento notable en la última década. El número de publicaciones en el área de predicción de salud poblacional mediante aprendizaje automático aumentó de forma notable a partir de 2007, con la mayor representación de estudios procedentes de Estados Unidos y China (Morgenstern et al., 2020).

En el ámbito específico de la SAP, el trabajo de Clark et al. (2021) constituye el antecedente más directo del presente estudio. Aplicando tres algoritmos de aprendizaje automático —regresión logística regularizada, Random Forest y SVM— sobre datos de la encuesta BRFSS con 449.492 participantes adultos, los autores obtuvieron valores de AUC de entre 0.80 y 0.81, con factores socioeconómicos como ingresos y educación entre los predictores más relevantes. Este trabajo demuestra la viabilidad de predecir la SAP con buena capacidad discriminativa a partir de datos de encuesta, aunque se limita al contexto estadounidense y no restringe los predictores a determinantes conductuales.

En el contexto europeo, Gumà et al. (2023) aplicaron árboles de clasificación y Random Forest sobre datos de la encuesta SHARE con 27.230 participantes europeos de entre 65 y 79 años, confirmando la alta capacidad predictiva de las condiciones crónicas, las limitaciones funcionales y la depresión sobre la SAP. Asimismo, Loef et al. (2022) utilizaron Random Forest sobre datos del Estudio de Cohorte de Doetinchem de 30 años para identificar predictores longitudinales de la salud percibida, obteniendo un AUC de 0.71 e identificando que solo nueve exposiciones de distintos dominios eran responsables de la mayor parte del rendimiento del modelo.

Un antecedente metodológico de especial relevancia es el de Bozorgmehr y Weltermann (2023), que desarrollaron un modelo XGBoost interpretable para predecir niveles de estrés crónico a partir de datos de la encuesta nacional de salud alemana DEGS1, utilizando valores SHAP para identificar los factores protectores y de riesgo más relevantes. Este trabajo es metodológicamente muy próximo al presente, al combinar datos de encuesta nacional representativa, algoritmos de ensamble y técnicas de interpretabilidad basadas en SHAP.

2.2 Contexto y justificación

El presente trabajo se sitúa en la convergencia de tres áreas de conocimiento: la epidemiología de la salud percibida, la prevención del cáncer basada en población y la aplicación de técnicas de aprendizaje automático a datos de encuestas sanitarias nacionales.

Tomando el punto de vista epidemiológico, la evidencia acumulada establece que los principales factores de riesgo de cáncer —tabaco, alcohol, obesidad, sedentarismo y alimentación inadecuada— son exactamente los mismos determinantes conductuales que predicen con mayor consistencia una SAP deteriorada. La ESdE constituye un elemento básico de cohesión territorial para el seguimiento poblacional de las estrategias sanitarias del Sistema Nacional de Salud, y entre sus ámbitos de monitorización se encuentran explícitamente el tabaco, la obesidad, el alcohol y los factores de riesgo de cáncer. Desde lo metodológico, la fuente de datos empleada ofrece condiciones especialmente favorables para este tipo de análisis, pues muestra representatividad nacional y autonómica. El diseño probabilístico estratificado garantiza la representatividad de la muestra y permite la generalización de los resultados a la población adulta española no institucionalizada, lo que constituye una ventaja estructural frente a estudios basados en muestras de conveniencia o registros clínicos.

Respecto con la aportación al campo, el presente trabajo cubre una brecha específica identificada en la revisión de la literatura. Los estudios de predicción de SAP mediante aprendizaje automático existentes se han desarrollado principalmente en el contexto anglosajón —con la encuesta BRFSS estadounidense como fuente más utilizada— o sobre poblaciones europeas de edad avanzada con énfasis en morbilidad crónica (Gumà et al., 2023; Loef et al., 2022). No se ha identificado ningún estudio previo que aplique este enfoque sobre los microdatos de la encuesta nacional española con un modelo restringido explícitamente a determinantes conductuales y sociales, excluyendo la información clínica de morbilidad. Esta restricción, lejos de ser una limitación, es el núcleo de la contribución: permite cuantificar el peso predictivo de los hábitos de vida sobre la salud percibida de forma aislada, con implicaciones directas para el diseño de intervenciones preventivas en el ámbito del cáncer.

2.3 Planteamiento del problema

Lo anterior permite identificar la brecha que motiva el presente trabajo. La literatura epidemiológica ha establecido de forma robusta tanto la validez predictiva de la salud autopercibida como la relación entre los principales hábitos de vida y el riesgo de desarrollar cáncer. Por su parte, los estudios recientes de aprendizaje automático aplicados a encuestas de salud han demostrado que es posible predecir la SAP con buena capacidad discriminativa a partir de datos autodeclarados. Sin embargo, la intersección de estos tres cuerpos de evidencia (predicción de SAP, factores de riesgo oncológico y aprendizaje automático sobre encuesta nacional española) no ha sido abordada en la literatura disponible.

De forma específica, se detectan tres carencias que justifican este trabajo. En primer lugar, los modelos predictivos de SAP publicados incluyen habitualmente variables de morbilidad clínica como predictores, lo que genera una circularidad conceptual: se predice salud percibida usando información sobre enfermedades diagnosticadas, cuando el interés real es identificar individuos en riesgo antes de que la enfermedad se manifieste. En segundo lugar, ningún estudio previo identificado ha aplicado este enfoque sobre los microdatos de la ESdE con restricción explícita a variables conductuales y sociales. En tercer lugar, la interpretabilidad de los modelos no ha sido sistemáticamente abordada en el contexto español mediante técnicas como los valores SHAP.

Este trabajo plantea dar respuesta a estas carencias desarrollando un modelo de clasificación supervisada que prediga la salud percibida negativa —como indicador indirecto del riesgo percibido de desarrollar cáncer— a partir exclusivamente de hábitos autodeclarados y determinantes sociodemográficos recogidos en la ESdE 2023, e interpretando sus resultados mediante valores SHAP para cuantificar la contribución relativa de cada determinante. El resultado esperado es un modelo metodológicamente sólido, conceptualmente coherente con el marco de determinantes sociales de la salud, y con potencial de aplicación en la identificación de subgrupos de riesgo a escala poblacional.

Capítulo 3. OBJETIVOS

3.1 Objetivos generales

Desarrollar y evaluar un modelo de clasificación supervisada capaz de predecir la salud percibida negativa como indicador indirecto del riesgo percibido de desarrollar cáncer en la población adulta española, a partir de hábitos autodeclarados y determinantes sociodemográficos registrados en la Encuesta de Salud de España 2023, e interpretar sus resultados mediante técnicas de inteligencia artificial explicable para identificar los determinantes con mayor peso predictivo.

3.2 Objetivos específicos

- Integrar los cuestionarios de adultos y hogares de la ESdE 2023 en un único conjunto de datos analítico a nivel de persona, incorporando tanto determinantes individuales de salud como variables de contexto socioeconómico del hogar.
- Construir la variable dependiente binaria a partir de la pregunta de salud autopercibida de la ESdE 2023, aplicando el criterio de dicotomisación establecido por Eurostat para el Módulo Mínimo Europeo de Salud, que define como valoración positiva la suma de las categorías "bueno" y "muy bueno".
- Seleccionar un conjunto de variables predictoras representativas de los principales determinantes de salud modificables excluyendo aquellas que expresen el estado de salud actual o la morbilidad diagnosticada, para garantizar la validez conceptual del modelo como instrumento predictivo basado exclusivamente en hábitos y determinantes sociales.
- Diseñar y aplicar un pipeline de preprocesamiento adecuado a la naturaleza heterogénea de los datos de encuesta, que incluya estrategias diferenciadas de imputación, codificación y escalado según el tipo de cada variable, y que garantice la ausencia de fuga de información entre los conjuntos de entrenamiento y prueba.
- Comparar el rendimiento predictivo de tres algoritmos de clasificación mediante validación cruzada estratificada de cinco pliegues sobre el conjunto de entrenamiento, y evaluación final sobre un conjunto de prueba independiente, utilizando el AUCC-ROC como métrica principal de discriminación.
- Interpretar los resultados del modelo de mayor rendimiento mediante valores SHAP, cuantificando la contribución marginal de cada variable predictora, la dirección de sus efectos y la importancia agregada de cada bloque temático de determinantes.

3.3 Beneficios del proyecto

Se busca cubrir la brecha identificada en la literatura sobre la aplicación de aprendizaje automático a encuestas de salud con restricción explícita a determinantes conductuales. Los resultados del análisis proporcionan evidencia cuantitativa sobre la importancia relativa de los distintos bloques de hábitos como predictores de la salud percibida en la población española adulta, información que no estaba disponible de forma integrada en estudios previos.

Respecto a salud pública, el modelo desarrollado tiene potencial de aplicación en la identificación de subgrupos de riesgo a escala poblacional sin necesidad de datos clínicos, lo que lo hace replicable con cada nueva edición de la ESdE. La jerarquía de variables identificadas puede orientar el diseño de intervenciones preventivas priorizando los hábitos con mayor impacto sobre la salud percibida, con implicaciones directas para las estrategias de prevención del cáncer del Sistema Nacional de Salud.

Desde el punto de vista metodológico, el trabajo documenta de forma detallada un método reproducible de integración, preprocesamiento y modelado sobre microdatos de salud, que puede servir de referencia para investigaciones posteriores con la misma fuente de datos.

Capítulo 4. DESARROLLO DEL PROYECTO

4.1 Planificación del proyecto

El desarrollo del TFM se organizó en seis fases secuenciales, ejecutadas entre febrero y abril de 2026. La Tabla 1 corresponde al cronograma y duración aproximada de cada actividad.

- **Tarea 1: Revisión bibliográfica y estado del arte**

Se llevó a cabo una revisión sistemática de la literatura científica en torno a tres ejes temáticos: la salud autopercibida como indicador epidemiológico, las variables de salud modificables asociados al cáncer, y la aplicación de técnicas de aprendizaje automático a encuestas poblacionales de salud. Se consultaron bases de datos científicas internacionales como PubMed, Google Scholar y Scopus, así como fuentes institucionales del INE, el Ministerio de Sanidad, la SEOM, la OMS e INEC.

Las actividades principales desarrolladas fueron: búsqueda y selección de referencias bibliográficas relevantes, síntesis del estado del arte por bloques temáticos e identificación de la brecha metodológica que justifica el trabajo.

El hito alcanzado es: marco teórico y antecedentes metodológicos definidos.

- **Tarea 2: Obtención y exploración de los microdatos**

Se procedió a la obtención de los microdatos de la Encuesta de Salud de España 2023 desde el repositorio oficial del INE, disponibles de forma gratuita para fines de investigación. La exploración inicial abarcó los dos cuestionarios utilizados de adultos y hogares, así como sus respectivos diccionarios de variables, con el objetivo de comprender la estructura de la encuesta, el diseño del cuestionario y las posibilidades de integración entre ambas fuentes.

Las actividades principales desarrolladas fueron: descarga y carga de los microdatos en formato SPSS mediante Python, lectura de los diccionarios de variables, exploración de la distribución de la variable dependiente C1 e identificación preliminar de variables candidatas por bloque temático.

El hito alcanzado es: datos disponibles en el entorno de trabajo y variables candidatas identificadas.

- **Tarea 3: Integración, selección y preprocesamiento de variables**

Esta tarea constituyó el núcleo metodológico del trabajo y la más extensa en tiempo, debido a las decisiones iterativas sobre la idoneidad conceptual de cada variable. Se integraron los cuestionarios de adultos y hogares a nivel de persona, se aplicó un filtro automático de variables y se realizó una revisión teórica por bloques que llevó a excluir las variables de morbilidad clínica para garantizar la validez conceptual del modelo.

Las actividades principales desarrolladas fueron: integración de los cuestionarios mediante la variable IDENTHOGAR, filtro automático por missings, varianza y redundancia, revisión teórica por bloques temáticos con criterio de exclusión de variables de estado de salud actual,

tratamiento de missings estructurales derivados del flujo condicional del cuestionario, construcción de variables derivadas (IMC, sedentarismo en minutos) y diseño del pipeline de preprocesamiento diferenciado por tipo de variable.

El hito alcanzado es: dataset limpio, integrado y preprocesado con 57 variables predictoras finales.

- **Tarea 4: Modelado, entrenamiento y optimización de modelos**

Se entrenaron y compararon tres algoritmos de clasificación supervisada sobre el conjunto de entrenamiento: regresión logística, Random Forest y XGBoost. Cada modelo fue evaluado mediante validación cruzada estratificada de cinco pliegues y posteriormente sometido a una búsqueda de hiperparámetros mediante RandomizedSearchCV para mejorar su rendimiento.

Las actividades principales desarrolladas fueron: definición del pipeline completo de preprocesamiento con scikit-learn, entrenamiento y validación cruzada de los tres modelos base, búsqueda de hiperparámetros con RandomizedSearchCV y evaluación final de los modelos tuneados sobre el conjunto de prueba independiente.

El hito alcanzado es: tres modelos entrenados, optimizados y evaluados con métricas definitivas.

- **Tarea 5: Interpretación de resultados mediante SHAP y análisis de resultados**

Una vez identificado el modelo de mayor rendimiento, se calcularon los valores SHAP sobre el conjunto de prueba para interpretar la contribución de cada variable predictora a las predicciones del modelo. El análisis incluyó tanto la importancia global de las variables como la dirección de sus efectos y la importancia agregada por bloque temático.

Las actividades principales desarrolladas fueron: cálculo de valores SHAP con TreeExplainer sobre el conjunto de prueba, generación de visualizaciones, análisis comparativo entre el modelo restringido a determinantes conductuales y el modelo ampliado con variables de morbilidad, y síntesis de hallazgos por bloque temático.

El hito alcanzado es: análisis de interpretabilidad completo y jerarquía de variables cuantificada.

- **Tarea 6: Redacción de la memoria**

Se redactó la memoria del TFM siguiendo la estructura de la plantilla oficial de la Universidad Europea de Madrid, integrando los resultados obtenidos con el marco teórico desarrollado en la revisión bibliográfica.

Las actividades principales desarrolladas fueron: redacción de todos los capítulos de la memoria, elaboración de tablas y figuras de resultados e incorporación de referencias bibliográficas en formato APA 7.ª edición.

El hito alcanzado es: memoria del TFM completa y lista para entrega.

Actividades	Personas responsables e involucradas	Semana 1	Semana 2	Semana 3	Semana 4	Semana 5	Semana 6	Semana 7	Semana 8	Semana 9	Semana 10
Revisión bibliográfica y estado del arte	Investigador principal	X									
Obtención de data y exploración de esta	Investigador principal		X	X							
Integración de cuestionarios, selección y procesamiento de variables	Investigador principal				X	X					
Diseño del pipeline de modelado y entrenamiento de modelos	Investigador principal						X	X			
Evaluación, interpretación SHAP y análisis de resultados	Investigador principal								X		
Redacción de memoria y correcciones	Investigador principal									X	X

Tabla 4-1: Cronograma del TFM

4.2 Descripción de la solución, metodologías y herramientas empleadas

Todo el código desarrollado y descrito en esta sección se encuentra en la parte de Anexos.

4.2.1 Fuente de datos

La fuente de datos utilizada es la Encuesta de Salud de España 2023 (ESdE 2023), elaborada conjuntamente por el Instituto Nacional de Estadística y el Ministerio de Sanidad. La ESdE 2023 nace de la integración de la Encuesta Nacional de Salud (ENSE) y la Encuesta Europea de Salud de España (EESE) en una única operación estadística, enmarcada en el Plan Estadístico Nacional 2025-2028 y el Sistema de Información del SNS. Recoge información sanitaria relativa a la población residente en España en 24.673 hogares, mediante tres cuestionarios —hogar, adulto y menor— que abordan cuatro grandes áreas: sociodemográfica, estado de salud, utilización de los servicios sanitarios y determinantes de la salud (INE & Ministerio de Sanidad, 2025).

Para el presente trabajo se utilizaron exclusivamente los cuestionarios de adultos y hogares, que comprenden 434 y 56 variables respectivamente. El cuestionario de adultos recoge información individual sobre estado de salud, hábitos de vida, uso de servicios sanitarios y prácticas preventivas de la persona seleccionada en cada hogar. El cuestionario de hogares contiene información sobre composición del hogar, características de la vivienda, ingresos y clase social de la persona de referencia. Ambos cuestionarios se enlazan mediante el identificador de hogar (IDENTHOGAR), que permite la integración a nivel de persona.

4.2.2 Variable por predecir o dependiente

La variable dependiente se construyó a partir de la pregunta C1 del cuestionario de adultos: "¿Cómo ha sido su estado de salud en los últimos doce meses?", con cinco categorías de respuesta: muy bueno, bueno, regular, malo y muy malo. Siguiendo el criterio de dicotomización establecido por Eurostat para el Módulo Mínimo Europeo de Salud (Eurostat, 2018), las categorías se agruparon en dos valores:

- **0 → Salud percibida positiva:** muy bueno + bueno (67,1% de la muestra)
- **1 → Salud percibida negativa:** regular + malo + muy malo (32,9% de la muestra)

Este criterio está ampliamente respaldado por la literatura epidemiológica internacional (Idler & Benyamini, 1997; Clark et al., 2021) y permite la comparabilidad directa con estudios previos basados en encuestas europeas armonizadas.

Se consideró también usar tres categorías (tricotomización) manteniendo "regular" como categoría intermedia; sin embargo, dado que las categorías positivas agrupan el 67,1% de la muestra frente al 32,9% restante, introducir una tercera clase hubiese agravado el desbalance en las categorías negativas y reducido la estabilidad del modelo.

4.2.3 Selección de variables predictoras

El proceso de selección de variables se realizó en dos fases sucesivas. En la primera fase se aplicó un filtro automático que eliminó: variables con más del 40% de valores ausentes, variables con varianza casi nula (una categoría con presencia superior al 95% de la muestra) y variables con redundancia estructural identificada. En la segunda fase se realizó una revisión teórica por bloques temáticos, guiada por el marco de determinantes sociales de la salud de Dahlgren y Whitehead (1991), que llevó a excluir adicionalmente todas las variables que expresan estado de salud actual o morbilidad diagnosticada —limitaciones funcionales, enfermedades crónicas, consumo de medicamentos, hospitalizaciones y variables derivadas de estado de salud mental— para evitar la circularidad conceptual descrita en el planteamiento del problema.

El grupo final de variables fueron 57, agrupadas en ocho bloques como muestra la Tabla 2.

Bloque	**Variables principales**	**Nr.**
Sociodemográfico	Edad, sexo, nivel estudios, estado civil	6
Actividad física	Tipo actividad laboral, frecuencia de ejercicio, días de deporte	8
Tabaco	Situación tabáquica actual, exfumador, uso de cigarrillos electrónicos	3
Alcohol	Frecuencia de consumo, consumo medio diario	4
Alimentación	Frecuencia de consumo de 9 grupos de alimentos	9
Antropometría	Altura, peso, IMC	3
Apoyo social	Personas de apoyo, interés de otros	3
Hogar y entorno	Ingresos, tipo de hogar, problemas de vivienda	8
Prácticas preventivas	Cribados oncológicos, vacuna de gripe	13

Tabla 4-2: Variables predictoras finales

4.2.4 Preprocesamiento

El preprocesamiento se diseñó de forma diferenciada según el tipo de cada variable: numérica, ordinal, binaria o nominal. Se implementó pipelines de scikit-learn para garantizar que todas las transformaciones se ajustan exclusivamente sobre el conjunto de entrenamiento y se aplican sobre el de prueba, evitando cualquier fuga de información.

Respecto al tratamiento de variables, este difiere entre modelos pues la regresión logística asume linealidad y requiere codificación one-hot para ordinales; los basados en árboles (Random Forest y XGBoost) explotan directamente el orden natural de las categorías ordinales

mediante OrdinalEncoder. Las variables numéricas continuas como edad, IMC, sedentarismo, consumo de alcohol, altura, peso y variables de actividad física se imputaron por la mediana y se estandarizaron mediante StandardScaler; las ordinales (frecuencias de consumo alimentario, escalas de actividad física, nivel de ingresos, nivel de estudios y otras escalas con orden natural) se imputaron por la moda. Para variables binarias y nominales se imputó por la moda, con tratamiento explícito de categorías estructuralmente ausentes mediante la asignación de una categoría "No aplica". Las variables de cribado oncológico femenino (mamografía, citología vaginal) se imputaron con la categoría "Es varón" para los hombres, cuya ausencia de respuesta es estructural por diseño del cuestionario. Cabe mencionar que todas estas estandarizaciones, codificaciones e imputaciones se hicieron con funciones propias de la librería scikit-learn.

La muestra se dividió en conjunto de entrenamiento (80%, n = 16.825) y conjunto de prueba (20%, n = 4.207) mediante muestreo estratificado por la variable dependiente, preservando la proporción 67% y 33% en ambos subconjuntos.

4.2.5 Modelos de clasificación

Se compararon tres algoritmos de clasificación supervisada con características complementarias, siguiendo el enfoque empleado en estudios análogos de predicción de salud poblacional (Clark et al., 2021; Bozorgmehr & Weltermann, 2023).

- **Regresión logística**

La regresión logística es un modelo estadístico paramétrico que estima la probabilidad de pertenencia a una clase binaria en función de un conjunto de variables predictoras. El modelo asume una relación lineal entre las variables independientes y el logaritmo de la razón de probabilidades (*log-odds*), expresada como:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n$$

Donde p es la probabilidad de que la observación pertenezca a la clase positiva (salud percibida negativa: 1), $\beta_0$es el término independiente y $\beta_i$son los coeficientes asociados a cada variable predictora x_i. La probabilidad predicha se obtiene aplicando la función sigmoide:

$$p = \frac{1}{1 + e^{-(\beta_0 + \sum_{i=0}^{n} \beta_i x_i)}}$$

Este modelo es el más interpretable, ya que los coeficientes $\beta_i$permiten cuantificar directamente el efecto de cada variable sobre el log-odds de la variable dependiente.

- **Random Forest**

Random Forest es un método de ensamble que construye un número determinado de árboles de decisión sobre submuestras aleatorias del conjunto de entrenamiento y promedia sus predicciones. Cada árbol se entrena sobre una muestra con reemplazo de los datos, y en cada nodo de división solo se considera un subconjunto aleatorio de variables, lo que introduce

diversidad entre los árboles y reduce la varianza del modelo. La predicción final para clasificación es la moda de las predicciones individuales de todos los árboles:

$$\hat{y} = moda(\widehat{y_1}, \widehat{y_2}, \cdots, \widehat{y_T})$$

donde T es el número de árboles. Los principales hiperparámetros configurados fueron: n_estimators (número de árboles), que controla la estabilidad de la estimación a mayor valor; max_depth (profundidad máxima de cada árbol), que regula la complejidad del modelo y previene el sobreajuste; max_features (fracción de variables consideradas en cada división), que controla la diversidad entre árboles; min_samples_leaf (número mínimo de observaciones en cada hoja), que actúa como regularización adicional; y max_samples (fracción de observaciones usadas en cada árbol), que controla el grado de bagging. El desbalance de clases se corrigió igualmente con class_weight='balanced'.

- **XGBoost**

XGBoost (eXtreme Gradient Boosting) es un método de ensamble basado en boosting con gradiente (Chen & Guestrin, 2016). A diferencia del bagging, el boosting construye los árboles de forma secuencial: cada árbol nuevo se entrena para corregir los errores residuales del conjunto anterior, minimizando una función de pérdida mediante descenso de gradiente. La función objetivo que XGBoost minimiza en cada iteración es:

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)\right) + \Omega(f_t)$$

donde $l$es la función de pérdida —log-loss para clasificación binaria—, $f_t$es el árbol añadido en la iteración t, y $\Omega(f_t)$es el término de regularización que penaliza la complejidad del árbol.

La configuración final empleó n_estimators = 300, que define el número de árboles en la secuencia; learning_rate = 0.05, tasa de aprendizaje ηque escala la contribución de cada árbol individual, favoreciendo la convergencia gradual y reduciendo el riesgo de sobreajuste; max_depth = 6, que limita la profundidad máxima de cada árbol controlando la complejidad del modelo; y scale_pos_weight, calculado como el cociente entre el número de observaciones de la clase negativa y la positiva:

$$\text{scale\_pos\_weight} = \frac{N_{\text{clase 0}}}{N_{\text{clase 1}}} = \frac{14.111}{6.921} \approx 2.04$$

Este parámetro corrige el desbalance de clases asignando mayor peso a la clase positiva (salud percibida negativa) durante el entrenamiento. La función de pérdida utilizada para la evaluación interna fue logloss.

Los tres modelos se evaluaron mediante validación cruzada estratificada de cinco pliegues sobre el conjunto de entrenamiento, preservando la proporción de clases en cada pliegue. El modelo final se entrenó sobre la totalidad del conjunto de entrenamiento y se evaluó sobre el conjunto de prueba independiente. Estas evaluaciones son validaciones internas, ambas usan datos de la misma encuesta y el mismo periodo; hacerlo sobre datos externos se reserva como futura línea de trabajo como se menciona en el Capítulo 7.

4.2.6 Métricas de evaluación

La evaluación de los modelos se realizó en dos etapas. Durante la validación cruzada sobre el conjunto de entrenamiento se calcularon tres métricas: AUC-ROC, F1-score ponderado y accuracy. En cambio, en el conjunto de test se calcularon AUC-ROC, AUC-PR, precisión, recall y F1-score. A continuación, se presenta una breve descripción de cada una.

- **AUC_ROC**

El área bajo la curva ROC mide la capacidad discriminativa del modelo con independencia del umbral de clasificación. La curva ROC representa la tasa de verdaderos positivos (sensibilidad) frente a la tasa de falsos positivos (1 − especificidad) para todos los umbrales posibles. Un AUC de 0.5 equivale a un clasificador aleatorio y un AUC de 1.0 a un clasificador perfecto:

$$\text{AUC-ROC} = \int_0^1 \text{TPR}(t)\, d\text{FPR}(t)$$

Es la métrica principal de comparación entre modelos por ser robusta al desbalance de clases y no depender de un umbral fijo de clasificación.

- **F1-score**

Es la media armónica de precisión y recall, calculada para cada clase y ponderada por el soporte (número de observaciones reales de cada clase), lo que la hace adecuada en presencia de desbalance:

$$F1_{\text{ponderado}} = \sum_k w_k \cdot 2 \cdot \frac{\text{Precisión}_k \times \text{Recall}_k}{\text{Precisión}_k + \text{Recall}_k}$$

donde $w_k$es la proporción de observaciones de la clase $k$sobre el total.

- **Accuracy**

La exactitud mide la proporción global de predicciones correctas sobre el total de observaciones:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Se reporta como métrica complementaria, aunque su interpretación es limitada en presencia de desbalance de clases, ya que un clasificador que predijera siempre la clase mayoritaria obtendría una accuracy del 67%.

- **AUC-PR**

El área bajo la curva precisión-recall complementa al AUC-ROC en escenarios de desbalance, evaluando el rendimiento del modelo específicamente sobre la clase positiva (salud percibida

negativa) a lo largo de todos los umbrales posibles. Es especialmente informativo cuando el interés se centra en la detección correcta de la clase minoritaria.

- **Recall**

Mide la proporción de casos de cada clase que el modelo identifica correctamente. Desde el punto de vista de la salud pública, el recall de la clase positiva es la métrica clínicamente más relevante, ya que un falso negativo —no identificar a un individuo con salud percibida negativa— tiene mayor coste social que un falso positivo:

$$\text{Recall} = \frac{TP}{TP + FN}$$

- **Precisión**

Esta mide la proporción de predicciones positivas que son efectivamente correctas:

$$\text{Precisión} = \frac{TP}{TP + FP}$$

4.2.7 Interpretabilidad de valores

Los valores SHAP (Shapley Additive explanations) son un método de interpretabilidad post-hoc fundamentado en la teoría de juegos cooperativos de Shapley (1953). Su principio central es asignar a cada variable predictora una contribución justa y consistente a cada predicción individual del modelo, respondiendo a la siguiente pregunta: ¿en qué medida cada variable aumenta o disminuye la predicción respecto a la predicción media del modelo?

Formalmente, el valor SHAP de la variable $j$para la observación $i$se define como:

$$\phi_j(i) = \sum_{S \subseteq F \setminus \{j\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} \left[f_{S \cup \{j\}}(x_i) - f_S(x_i) \right]$$

donde Fes el conjunto de todas las variables, Ses un subconjunto de variables que no incluye a j, y $f_S(x_i)$es la predicción del modelo usando únicamente las variables del subconjunto S. El valor SHAP es la contribución marginal promediada sobre todas las posibles coaliciones de variables.

Los valores SHAP satisfacen tres propiedades fundamentales que los hacen especialmente adecuados para interpretación en salud pública. La *eficiencia* garantiza que la suma de todos los valores SHAP de una observación más el valor base (predicción media del modelo) es igual a la predicción del modelo para esa observación:

$$f(x_i) = \phi_0 + \sum_{j=1}^{p} \phi_j\,(i)$$

La *simetría* garantiza que dos variables con la misma contribución marginal reciben el mismo valor SHAP. La *consistencia* garantiza que, si la contribución de una variable aumenta en el modelo, su valor SHAP nunca disminuye.

- **Implementación**

Para modelos basados en árboles como XGBoost y Random Forest, el TreeExplainer de la librería 'shap' calcula los valores de forma exacta y eficiente, aprovechando la estructura arbórea del modelo sin necesidad de aproximaciones (Lundberg & Lee, 2017). Los valores SHAP se calcularon sobre el conjunto de prueba completo (n = 4.207), lo que permite obtener estimaciones representativas de la población.

Un valor SHAP positivo para una variable en una observación concreta indica que dicha variable empujó la predicción hacia la clase 1 (salud percibida negativa), mientras que un valor SHAP negativo indica que empujó hacia la clase 0 (salud positiva). La magnitud del valor indica la intensidad de esa contribución.

4.3 Recursos requeridos

Para la ejecución del presente proyecto se emplearon los siguientes recursos listados:

Datos:

- Microdatos de la Encuesta de Salud de España 2023 (cuestionarios de adultos y hogares), obtenidos gratuitamente del repositorio oficial del Instituto Nacional de Estadística.
- Diccionarios de variables y documentación metodológica de la ESdE 2023, publicados por el INE y el Ministerio de Sanidad.

Software:

- Python 3.12 como lenguaje de programación principal.
- Librerías: pandas, numpy, pyreadstat, scikit-learn, xgboost, shap, matplotlib y seaborn.
- Jupyter Notebook como entorno de desarrollo interactivo.
- Microsoft Word para la redacción de la memoria.

Hardware:

- Ordenador personal con procesador Intel Core i7, 16 GB de RAM y sistema operativo Windows 11. No se requirió infraestructura de computación en la nube dado el tamaño manejable del conjunto de datos.

Recursos humanos:

- Autor del trabajo: dedicación aproximada de 300 horas distribuidas a lo largo de las seis fases del proyecto.

- Director del TFM: supervisión y orientación metodológica a lo largo del desarrollo del trabajo.

Bibliografía:

- Acceso a bases de datos científicas: PubMed, Google Scholar y Scopus.
- Publicaciones institucionales de la OMS, SEOM, REDECAN, INE y Ministerio de Sanidad.

4.4 Presupuesto

El presupuesto recoge la evaluación económica total del proyecto, incluyendo el coste del tiempo dedicado por el autor, los recursos técnicos empleados y el software utilizado. Todos los recursos enumerados en la sección 4.3 están recogidos en la Tabla 3.

Tipo de coste	Valor	Comentarios
Horas dedicadas por el autor	$ 7500	300 horas a 25 $/hora, distribuidas en las seis fases del proyecto
Equipo informático	$ 1000	Ordenador personal, valor de mercado aproximado
Lenguaje de programación	$ 0	Software de código abierto sin coste de licencia
Software de ofimática	$ 150	Licencia Microsoft 365 Personal (precio anual de referencia)
Bases de datos	$ 0	Datos públicos de acceso gratuito

Tabla 4-3: Presupuestos del proyecto

Los costes de software y datos son nulos gracias al uso de herramientas de código abierto y fuentes de datos públicas, lo que pone de manifiesto la viabilidad económica de este tipo de investigaciones basadas en datos abiertos y software libre.

4.5 Viabilidad

El coste total del proyecto asciende a 8.650 dólares, siendo el tiempo del autor el componente mayoritario. Frente a este coste, los beneficios potenciales del trabajo operan en dos dimensiones. En el ámbito científico, el modelo y la metodología desarrollados son reutilizables con cada nueva edición de la ESdE, que tiene periodicidad trienal, lo que multiplica el valor del

trabajo sin coste adicional significativo. En el ámbito de la salud pública, si el modelo contribuyera a orientar intervenciones preventivas focalizadas en los subgrupos de mayor riesgo, el ahorro potencial para el Sistema Nacional de Salud superaría con creces el coste del proyecto, por lo que incluso una mejora marginal en la detección temprana o en la focalización de programas preventivos tendría un retorno económico positivo.

Respecto a la parte técnica, el proyecto es viable y reproducible sin infraestructura especializada. Todo el análisis se ejecutó sobre un ordenador personal estándar en un tiempo razonable, utilizando exclusivamente herramientas de código abierto y datos públicos. El pipeline desarrollado en Python puede ejecutarse íntegramente sobre los microdatos de futuras encuestas, lo que garantiza su sostenibilidad técnica a largo plazo.

4.6 Resultados del proyecto

4.6.1 Rendimiento de los modelos

Los tres modelos de clasificación fueron entrenados y evaluados con un conjunto de variables predictoras correspondientes a hábitos de vida, características sociodemográficas y contexto del hogar. La Tabla 4 recoge los resultados de la validación cruzada y la evaluación final sobre el conjunto de prueba independiente.

Modelo	CV AUC ± Test	AUCPR	AUC
Regresión Logística	0.79 ± 0.006	0.79	0.67
Random Forest	0.78 ± 0.007	0.78	0.66
XGBoost	0.79 ± 0.007	0.79	0.68

Tabla 4-4: Resultados comparativos de los modelos

Los tres modelos muestran un rendimiento similar, con AUC-ROC en torno a 0.79 y una brecha mínima entre la validación cruzada y la evaluación en test (inferior a 0.01 en los tres casos) lo que confirma la ausencia de sobreajuste y la buena capacidad de generalización de las soluciones desarrolladas. Además, en la Tabla 5 se presentan los resultados por clase sobre el conjunto de prueba para cada modelo. El recall de la clase positiva (salud negativa) oscila entre 0.66 y 0.72 en los tres modelos, lo que indica que entre dos tercios y tres cuartos de los individuos con salud percibida negativa son identificados correctamente a partir exclusivamente de sus hábitos y determinantes sociales. XGBoost obtiene el mejor equilibrio entre precisión y recall para la clase de interés, con un F1-score de 0.62.

En este caso XGBoost obtiene el mejor rendimiento en todas las métricas, seguido de regresión logística y Random Forest, por ello los resultados futuros mencionados serán con este modelo mencionado.

Modelo	Clase	Precisión	Recall	F1-score	Soporte
Regresión Logística	Salud positiva (0)	0.837	0.716	0.772	2,822
	Salud negativa (1)	0.553	0.716	0.624	1,385
	Accuracy global			0.716	4,207
Random Forest	Salud positiva (0)	0.817	0.747	0.780	2,822
	Salud negativa (1)	0.561	0.658	0.606	1,385
	Accuracy global			0.718	4,207
XGBoost	Salud positiva (0)	0.824	0.763	0.792	2,822
	Salud negativa (1)	0.580	0.668	0.621	1,385
	Accuracy global			0.731	4,207

Tabla 4-5: Resultado por modelo en conjunto de prueba

4.6.2 Importancia de variables

El análisis de interpretabilidad se realizó mediante valores SHAP calculados sobre el conjunto de prueba completo. La Figura 1 muestra un Bar Plot de importancia global de las 20 variables con mayor contribución predictiva.

La edad (EDADA) es con diferencia el predictor más potente del modelo, con un valor SHAP medio absoluto de 0.404, casi el doble que la segunda variable en el ranking. Entre las variables conductuales, el tiempo dedicado a caminar para desplazarse (O4) ocupa la segunda posición global (0.210), por encima del tipo de actividad física laboral (O1, 0.170) y el consumo de alcohol (R1, 0.169). La ausencia de cribado de sangre oculta en heces (L10_No, 0.123) figura en quinta posición, siendo el predictor preventivo más relevante del modelo. El sedentarismo en minutos diarios (0.120), el IMC calculado (0.120) y el apoyo social percibido (S1, 0.119) presentan importancias similares. El tabaco aparece en décima posición, representado por la categoría de no haber fumado nunca (Q1, 0.101), con efecto protector sobre la salud percibida.

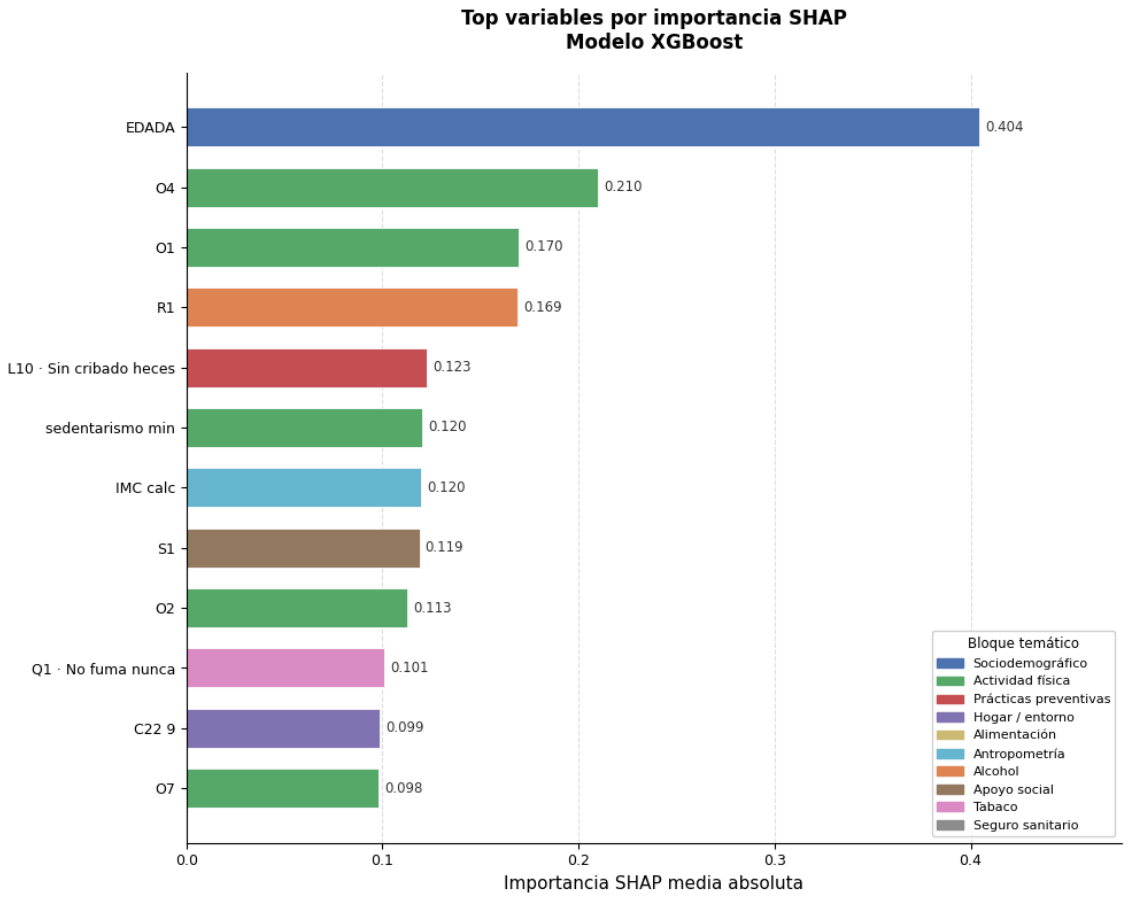

Figura 1: Top variables por importancia (SHAP)

Con el fin de cuantificar la contribución relativa de cada grupo de variables sobre la salud percibida, se calculó la importancia agregada por bloque temático como la suma de los valores SHAP medios absolutos de todas las variables de cada bloque. La Figura 2 muestra los resultados de lo antes mencionado. El bloque sociodemográfico acumula la mayor importancia agregada, impulsada principalmente por la edad, que por sí sola representa el 45.9% del total de ese bloque. La actividad física es el segundo bloque más relevante, con una distribución más equilibrada entre sus variables como: tiempo caminando, tipo de actividad laboral, frecuencia de ejercicio y sedentarismo, lo que sugiere que el patrón global de actividad física tiene mayor peso predictivo que cualquier indicador individual. Las prácticas preventivas ocupan la tercera posición, con los cribados oncológicos como variables más destacadas dentro del bloque. El bloque de hogar y entorno refleja el gradiente socioeconómico en salud, con las condiciones del entorno residencial como determinante relevante. La alimentación, la antropometría, el alcohol y el apoyo social presentan importancias similares. El tabaco, pese a ser uno de los factores de riesgo oncológico más documentados, ocupa una posición intermedia en el ranking, lo que puede explicarse por la reducción de la prevalencia de tabaquismo en la muestra y por la mediación de su efecto a través de variables de estado de salud que fueron excluidas del modelo.

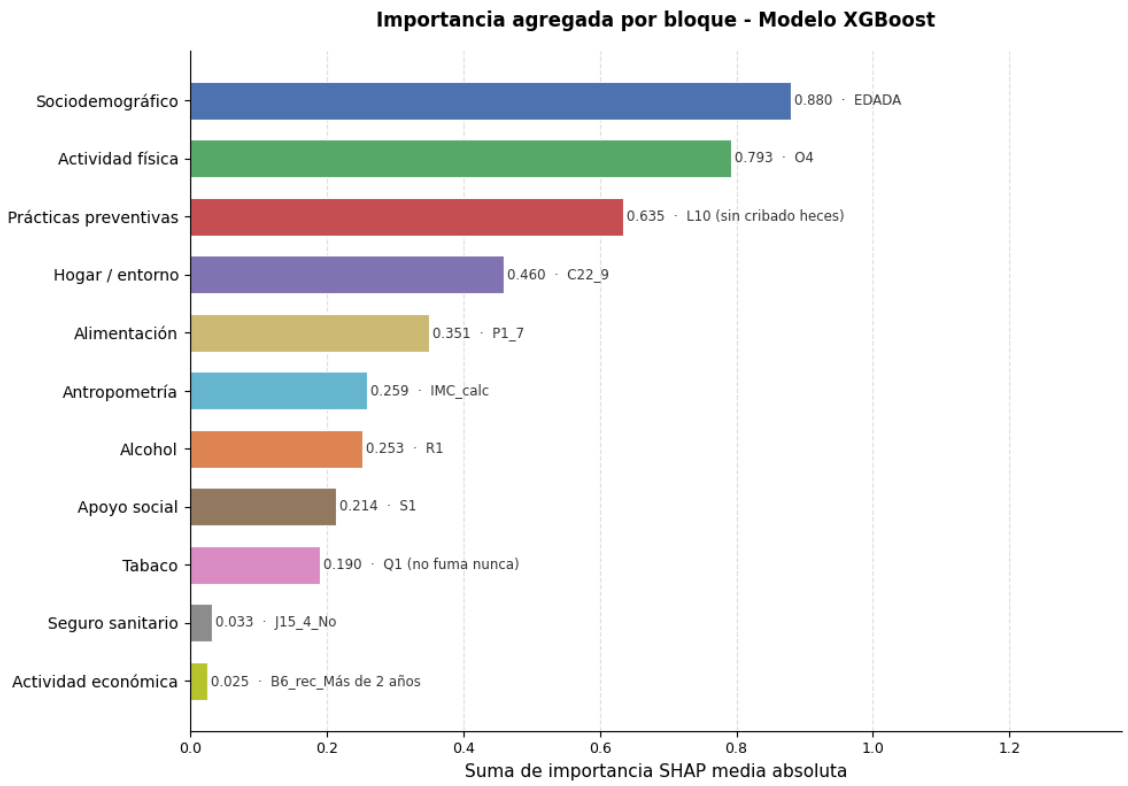

Figura 2: Importancia por bloque temático

Capítulo 5. DISCUSIÓN

Con todo lo anterior mencionado, se analiza los resultados obtenidos en relación con la literatura existente, reflexiona sobre las decisiones metodológicas adoptadas a lo largo del proyecto y discute las implicaciones de los hallazgos. Se abordan asimismo las limitaciones inherentes al diseño del estudio y a la fuente de datos utilizada, con el fin de contextualizar adecuadamente el alcance de las conclusiones.

5.1 Resultados

A continuación, se presentan los resultados obtenidos y los cuales deben interpretarse como asociaciones predictivas y no relaciones causales, dado el diseño transversal del proyecto. Las variables que detectamos en el modelo son predictoras de la salud autopercibida negativa mas no causas.

El AUC-ROC de 0.79 obtenido por XGBoost en el conjunto de prueba es consistente con los valores reportados en estudios comparables de predicción de SAP mediante aprendizaje automático sobre datos de encuesta poblacional. Clark et al. (2021) obtuvieron valores de AUC de entre 0.80 y 0.81 aplicando regresión logística regularizada, Random Forest y SVM sobre la encuesta BRFSS con 449.492 participantes, aunque su modelo incluía variables de morbilidad clínica que el presente trabajo excluye explícitamente. Loef et al. (2022) obtuvieron un AUC de 0.71 usando Random Forest sobre datos del Estudio de Cohorte de Doetinchem con exposiciones del ciclo vital, en un contexto longitudinal más restrictivo. El resultado de este trabajo se sitúa entre ambos referentes, lo que sugiere que la capacidad predictiva de los determinantes conductuales y sociales sobre la SAP es elevada y equiparable a la de modelos más completos cuando se aplica sobre una encuesta nacional representativa.

Uno de los hallazgos más importantes es y menos esperado del análisis es la posición de la variable O4 (tiempo dedicado a caminar para desplazarse) como segundo predictor global del modelo, por encima del tipo de actividad laboral, la frecuencia de ejercicio en tiempo libre y los días de práctica deportiva. Este resultado indica que la actividad física integrada en la vida cotidiana como forma de desplazamiento tiene mayor peso predictivo sobre la SAP que el ejercicio estructurado en tiempo de ocio. Este hallazgo es consistente con evidencia reciente que destaca el papel protector del caminar como forma de actividad física cotidiana. Estudios sobre movilidad activa han documentado que el desplazamiento a pie se asocia con mejor SAP y menor IMC, independientemente de otros indicadores de actividad física (Lindström et al., 2016). Con esto, las políticas que favorecen entornos caminables y el uso de la movilidad activa como forma de desplazamiento habitual pueden tener un impacto sobre la salud percibida superior al de los programas de promoción del ejercicio deportivo formal.

El consumo de alcohol (R1) se sitúa en cuarta posición global mientras que el tabaco (Q1) aparece en décima posición. Esto puede parecer contraintuitivo dado que el tabaco es el principal factor de riesgo oncológico documentado, pero tiene una explicación metodológica coherente: en la ESdE 2023, el 16,6% de la población de 15 y más años afirma fumar a diario, lo

que supone 3,2 puntos menos que en 2020 (INE & Ministerio de Sanidad, 2025). La reducción de la prevalencia de tabaquismo en la muestra limita la variabilidad de la variable y, por tanto, su capacidad discriminativa en el modelo. El alcohol, con una prevalencia mucho mayor en la población española, posee mayor variabilidad y genera una señal predictiva más grande.

5.2 Limitaciones

Este proyecto presenta limitaciones que deben tenerse en cuenta al momento de interpretar resultados y valorar el alcance de las conclusiones.

- **Diseño transversal**

La principal limitación estructural del estudio es su naturaleza transversal: tanto la variable dependiente como todas las variables predictoras se midieron en el mismo momento temporal. Esto impide establecer relaciones causales entre los hábitos de vida y la salud percibida. No es posible determinar, con los datos disponibles, si los hábitos preceden y determinan la SAP o si, por el contrario, un estado de salud deteriorado lleva a modificar los hábitos. Para establecer causalidad serían necesarios datos longitudinales con mediciones repetidas en el tiempo, como los empleados por Loef et al. (2022) en el Estudio de Cohorte de Doetinchem. Los resultados del presente trabajo deben interpretarse, por tanto, como asociaciones predictivas y no como relaciones causales.

- **Sesgo social en variables autodeclaradas**

Las variables predictoras son autodeclaradas, lo que introduce un potencial sesgo de deseabilidad social. El consumo de alcohol y tabaco, en particular, tienden a subestimarse en encuestas de salud debido a la naturaleza de estos hábitos. Este sesgo es estructural en cualquier encuesta de salud poblacional y no es exclusivo de la ESdE, pero limita la precisión de las estimaciones asociadas a estos predictores.

- **Restricción a población no institucionalizada**

La ESdE 2023 se dirige exclusivamente a la población residente en viviendas familiares, excluyendo a las personas institucionalizadas como residencias de mayores, centros de larga estancia, hospitales de crónicos. Esta población, que presenta en general un peor estado de salud percibida y una mayor exposición acumulada a factores de riesgo, no está representada en el modelo. Los resultados son, por tanto, generalizables únicamente a la población adulta española no institucionalizada.

- **Ausencia de validación externa**

El modelo fue entrenado y evaluado exclusivamente sobre datos de la ESdE 2023. No se realizó una validación externa sobre datos de una encuesta diferente o de una edición anterior de la misma encuesta, lo que limita la evaluación de la generalización del modelo más allá de la muestra disponible. La validación cruzada estratificada proporciona una estimación robusta del rendimiento interno, pero no sustituye a una validación sobre datos completamente independientes.

5.3 Reflexiones

El título del trabajo —*Modelado de riesgo percibido de desarrollar cáncer según hábitos autodeclarados*— planteó desde el inicio un reto metodológico significativo: la Encuesta de Salud de España 2023 no contiene ninguna pregunta que mida directamente la percepción de riesgo oncológico por parte del encuestado. Ante esta realidad, se evaluaron tres aproximaciones posibles:

- La primera consistía en cambiar el título para alinearlo con la variable disponible.
- La segunda, en construir una variable compuesta a partir de indicadores de diagnóstico de tumor maligno (C5a_27) o de comportamiento preventivo oncológico.
- La tercera, en mantener el título y justificar el uso de la salud autopercibida como indicador indirecto del riesgo percibido de desarrollar cáncer, apoyándose en la literatura prospectiva que documenta que la SAP deteriorada precede al diagnóstico oncológico.

Se optó por la tercera vía por ser la más sólida desde el punto de vista científico y la más coherente con la evidencia disponible. La revisión bibliográfica confirmó que los factores de riesgo de cáncer más documentados son exactamente los mismos determinantes que predicen con mayor consistencia una SAP negativa, lo que valida conceptualmente el puente establecido entre ambos constructos.

Respecto a la parte técnica del proyecto, el proceso de selección de variables fue sustancialmente más complejo de lo inicialmente previsto. Cada iteración del análisis SHAP revelaba nuevos problemas: categorías vacías derivadas de la codificación de variables condicionales, variables de acceso sanitario con interpretación ambigua entre determinante y consecuencia de salud, y variables laborales con alta tasa de vacíos estructural. Este proceso de depuración no estaba contemplado con esa extensión en el planteamiento inicial del trabajo. Sin embargo, esta complejidad aportó valor metodológico al resultado final: el modelo definitivo es el producto de un proceso explícito y documentado de toma de decisiones sobre la idoneidad conceptual de cada variable, lo que añade transparencia y reproducibilidad al análisis.

Capítulo 6. CONCLUSIONES

6.1 Conclusiones del trabajo

Las conclusiones que se obtuvieron están redactadas acorde a los objetivos generales y específicos que se plantearon en un principio. Con ello se presenta lo logrado respecto al objetivo general:

- Desarrollar y evaluar un modelo de clasificación supervisada capaz de predecir la salud percibida negativa como indicador indirecto del riesgo percibido de desarrollar cáncer en la población adulta española, a partir exclusivamente de hábitos autodeclarados y determinantes sociodemográficos y tener una interpretación de variables con cuantía de cuanto aporta cada una.

Para los objetivos específicos se logró:

- Integrar con éxito el cuestionario de adultos con el cuestionario de hogares, obteniendo un conjunto de datos analítico a nivel de persona que incorpora tanto determinantes individuales de salud como variables de contexto socioeconómico del hogar.
- Construir la variable dependiente binaria a partir de la pregunta de salud autopercibida, aplicando el criterio de dicotomización de Eurostat para el Módulo Mínimo Europeo de Salud.
- Seleccionar variables predictoras que garantizan la validez conceptual del modelo como instrumento predictivo basado exclusivamente en determinantes conductuales y sociales.
- Implementar un preprocesamiento diferenciado por tipo de variable: numéricas continuas, ordinales, binarias y nominales.
- Evaluar tres modelos, regresión logística, Random Forest y XGBoost, y obtener el mejor de ellos.
- Revelar qué predictores son los más potentes y cuáles resultaron tener menor importancia a lo que se pensaba.

6.2 Conclusiones personales

El desarrollo de este trabajo ha supuesto un recorrido que va mucho más allá del aprendizaje técnico con modelos ML, sino la investigación en salud pública. Las decisiones metodológicas más importantes no son las algorítmicas sino las conceptuales: qué variable mide realmente el fenómeno de interés, qué variables introducen circularidad en el modelo, dónde está el límite entre determinante y consecuencia. Estas preguntas no tienen respuesta en la documentación de librerías Python ni en los tutoriales de cómo usar un modelo; requieren conocimiento del dominio, lectura de literatura epidemiológica y disposición a replantear el enfoque cuando los resultados lo indican.

Desde el punto de vista personal, trabajar con los microdatos de la Encuesta de Salud de España ha sido una experiencia que conecta el trabajo técnico con una realidad concreta: los datos

representan a personas reales, con hábitos, condiciones de vida y percepciones de salud que son el resultado de realidades sociales que van mucho más allá de las decisiones individuales. Esa conexión entre los números y la realidad poblacional es lo que da sentido a este tipo de investigación y lo que espero que este trabajo, de una u otra manera, contribuya a iluminar.

Capítulo 7. FUTURAS LÍNEAS DE TRABAJO

La principal extensión del trabajo sería la validación longitudinal del modelo, esto permitiría evaluar si los hábitos autodeclarados en un momento temporal predicen cambios en la SAP en la siguiente medición, añadiendo validez predictiva prospectiva y fortaleciendo el vínculo entre hábitos y riesgo oncológico percibido. En la misma línea, el pipeline desarrollado es directamente reutilizable sobre microdatos de ediciones futuras de la encuesta, lo que garantiza la sostenibilidad del trabajo sin coste metodológico adicional.

Desde el punto de vista del modelado, resultaría de interés explorar modelos estratificados por grupos de edad, sexo o comunidad autónoma, dado que la heterogeneidad territorial en España en prevalencia de hábitos y condiciones socioeconómicas hace especialmente pertinente un análisis regional. Asimismo, la exploración de modelos de clasificación ordinal que mantengan las cinco categorías originales (muy bueno, bueno, regular, malo y muy malo) permitiría capturar gradientes más finos en la percepción de salud y evaluar si los determinantes identificados tienen efectos diferenciales a lo largo de toda la escala de SAP.

Finalmente, la integración con variables ambientales disponibles en fuentes externas enlazables por comunidad autónoma y estrato municipal, y la replicación del pipeline sobre encuestas equivalentes de otros países de la Unión Europea armonizadas con la EHIS de Eurostat, constituyen dos extensiones naturales del trabajo con potencial de contribución a la literatura comparativa europea sobre determinantes de la salud percibida.

Capítulo 8. REFERENCIAS

Benyamini, Y., & Idler, E. L. (1999). Community studies reporting association between self-rated health and mortality. *Research on Aging*, 21(3), 392–401. https://doi.org/10.1177/0164027599213002

Bozorgmehr, A., & Weltermann, B. (2023). Prediction of chronic stress and protective factors in adults: Development of an interpretable prediction model based on XGBoost and SHAP using national cross-sectional DEGS1 data. *JMIR AI*, 2, e41868. https://doi.org/10.2196/41868

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. https://doi.org/10.1145/2939672.2939785

Clark, C. R., Ommerborn, M. J., Moran, K., Brooks, K., Haas, J., Bates, D. W., & Wright, A. (2021). Predicting self-rated health across the life course: Health equity insights from machine learning models. *Journal of General Internal Medicine*, 36(5), 1181–1188. https://doi.org/10.1007/s11606-020-06438-1

Dahlgren, G., & Whitehead, M. (1991). *Policies and strategies to promote social equity in health*. Institute for Futures Studies.

Eurostat. (2018). *European Health Interview Survey (EHIS wave 3) — Methodological manual*. Publications Office of the European Union. https://ec.europa.eu/eurostat/web/products-manuals-and-guidelines/-/KS-02-18-240

Fink, H., Langselius, O., Vignat, J., Rumgay, H., Rehm, J., Martinez, R. X., Santero, M., Lopez-Perez, L., Inoue, M., Zeng, H., Shield, K., Morgan, E., Ilbawi, A., & Soerjomataram, I. (2026). Global and regional cancer burden attributable to modifiable risk factors to inform prevention. *Nature Medicine*. https://doi.org/10.1038/s41591-026-04219-7

Gumà-Lao, J., & Arpino, B. (2023). A machine learning approach to determine the influence of specific health conditions on self-rated health across education groups. *BMC Public Health*, 23, 131. https://doi.org/10.1186/s12889-023-15053-8

Idler, E. L., & Benyamini, Y. (1997). Self-rated health and mortality: A review of twenty-seven community studies. *Journal of Health and Social Behavior*, 38(1), 21–37. https://doi.org/10.2307/2955359

Instituto Nacional de Estadística. (2025). *Estadística de defunciones según la causa de muerte 2023*. https://www.ine.es/dyngs/INEbase/es/operacion.htm?c=Estadistica_C&cid=1254736176780&menu=ultiDatos&idp=1254735573175

Instituto Nacional de Estadística & Ministerio de Sanidad. (2025). *Encuesta de Salud de España 2023: resumen metodológico*. https://www.sanidad.gob.es/estadEstudios/estadisticas/encuestaSaludEspana/ESdE2023/ESdE2023_presentacion_web.pdf

Lindström, M., Hanson, B. S., & Östergren, P. O. (2016). Active traveling and its associations with self-rated health, BMI and physical activity: A comparative study in the adult Swedish population. *European Journal of Public Health*, 26(3), 208–213. https://doi.org/10.1093/eurpub/ckw005

Loef, B., Wong, A., Janssen, N. A. H., Strak, M., Hoekstra, J., Picavet, H. S. J., Boshuizen, H. C. H., Verschuren, W. M. M., & Herber, G. C. M. (2022). Using random forest to identify longitudinal predictors of health in a 30-year cohort study. *Scientific Reports*, 12, 10372. https://doi.org/10.1038/s41598-022-14632-w

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30, 4768–4777.

Ministerio de Sanidad. (2017). *Informe de la Encuesta Europea de Salud en España 2014*. Ministerio de Sanidad, Servicios Sociales e Igualdad.

Morgenstern, J. D., Buajitti, E., O'Neill, M., Piggott, T., Goel, V., Fridman, D., Kornas, K., & Rosella, L. C. (2020). Predicting population health with machine learning: A scoping review. *BMJ Open*, 10(10), e037860. https://doi.org/10.1136/bmjopen-2020-037860

Shapley, L. S. (1953). A value for n-person games. En H. Kuhn & A. Tucker (Eds.), *Contributions to the theory of games II* (pp. 307–317). Princeton University Press.

Sociedad Española de Oncología Médica & Red Española de Registros de Cáncer. (2026). *Las cifras del cáncer en España 2026*. SEOM.

Sociedad Española de Oncología Médica. (2024). *El ejercicio físico reduce hasta un 30% el riesgo de muchos cánceres*. https://seom.org/otros-servicios/noticias/210352-el-ejercicio-fisico-reduce-hasta-un-30-el-riesgo-de-muchos-canceres

Capítulo 9. ANEXOS

Anexo A. Código Python implementado

El código es extenso y por tanto rompería la estética el plasmarlo aquí. Por tanto, ha sido compartido en el siguiente repositorio de Github:

https://github.com/brackpug/implementacion_TFM.git

[PÁGINA INTENCIONADAMENTE EN BLANCO]