COMILLAS
UNIVERSIDAD PONTIFICIA

Facultad de Ciencias Económicas y Empresariales

ANÁLISIS DE LA DEPRESIÓN EN ESPAÑA MEDIANTE TÉCNICAS DE *MACHINE LEARNING*

Autor: Carolina Rivera Marassa
Director: Luis Ángel Calvo Pascual

MADRID | Junio 2026

ÍNDICE

CAPÍTULO I: INTRODUCCIÓN 8

1. CONTEXTO Y RELEVANCIA DE ESTUDIO 8
2. ANTECEDENTES 11
3. OBJETIVOS 12
4. METODOLOGÍA 14
5. ESTRUCTURA 15

CAPÍTULO II: TRATAMIENTO DE LOS DATOS 16

1. BASE DE DATOS UTILIZADA 16
2. PRESELECCIÓN DE VARIABLES 16
3. ANÁLISIS DESCRIPTIVO 18
   - **3.1 Análisis de las variables categóricas** 18
   - **3.2 Análisis de las variables numéricas** 22
   - **3.3 Relación entre la depresión y variables explicativas** 23
4. IDENTIFICACIÓN Y TRATAMIENTO DE VALORES ATÍPICOS 29
5. TRATAMIENTO DE VALORES PERDIDOS 30
   - **5.1 Imputación simple** 31
   - **5.2 Imputación múltiple** 32

CAPÍTULO III: SELECCIÓN DE VARIABLES 34

1. EXPLICACIÓN Y JUSTIFICACIÓN DEL USO DE *MUTUAL INFORMATION* 34
2. RESULTADOS Y SELECCIÓN DE VARIABLES 35

CAPÍTULO IV: MODELOS DE MACHINE LEARNING 37

1. MODELOS APLICADOS 38
   - **1.1 Regresión Logística** 38
   - **1.2 *XGBoost*** 38
2. COMPARACIÓN DE RESULTADOS 39

CAPÍTULO V: MODELO INTERPRETABLE 42

CAPÍTULO VI: REGLAS DE ASOCIACIÓN 47

1. METODOLOGÍA Y MÉTRICAS..... 47
2. RESULTADOS: REGLAS DE ASOCIACIÓN PARA DEPRESIÓN SEVERA 48
3. CONTRASTE VISUAL DE LOS FACTORES DE RIESGO..... 51

CAPÍTULO VII: CONCLUSIONES Y FUTURAS LÍNEAS DE INVESTIGACIÓN 54

1. CONCLUSIONES..... 54
2. LIMITACIONES..... 57
3. FUTURAS LÍNEAS DE INVESTIGACIÓN 58

DECLARACIÓN DE USO DE HERRAMIENTAS DE INTELIGENCIA ARTIFICIAL GENERATIVA 60

BIBLIOGRAFÍA 61

Índice de gráficos

Gráfico 1: Variación porcentual de cuadros depresivos severos activos por grupo etario en España (2020 y 2023) 8
Gráfico 2: Número de casos de depresión o trastornos depresivos registrados en España de 2011 a 2023 9
Gráfico 3: Distribución inicial de la variable objetivo depresión antes de la recodificación 21
Gráfico 4: Distribución de la variable objetivo depresión tras la recodificación 22
Gráfico 5: Edad según presencia de depresión 23
Gráfico 6: Porcentaje de depresión por grupo de edad 24
Gráfico 7: Prevalencia de depresión por apoyo social percibido 25
Gráfico 8: Prevalencia de depresión por convivencia en pareja 25
Gráfico 9: Prevalencia de depresión por estado civil 26
Gráfico 10: Porcentaje de depresión por Comunidad Autónoma 28
Gráfico 11: Importancia de las variables explicativas de la severidad depresiva según información mutua 36
Gráfico 12: Comparativa de métricas de evaluación por modelo 40
Gráfico 13: Curva ROC comparativa entre la regresión logística y XGBoost 40
Gráfico 14: Importancia global de las variables según SHAP — XGBoost 43
Gráfico 15: Diagrama de dispersión SHAP (beeswarm) 43
Gráfico 16: SHAP waterfall — perfil de alto riesgo (Comunidad Valenciana) 45
Gráfico 17: SHAP waterfall — perfil de bajo riesgo (Ceuta) 46
Gráfico 18: Diagrama de dispersión de las reglas de asociación para depresión severa según soporte, confianza y lift 49
Gráfico 19: Reglas de asociación con mayor lift para depresión severa 50
Gráfico 20: Sedentario sin tranquilizantes vs Sedentario con tranquilizantes 51
Gráfico 21: Hombre sedentario vs Mujer sedentaria 52
Gráfico 22: Hombre + tranquilizantes vs Mujer + tranquilizantes 52
Gráfico 23: Hombre mayor de 55 años vs Mujer mayor de 55 años 53

Índice de tablas

Tabla 1: Variables preseleccionadas y su relevancia teórica 17
Tabla 2: Resumen del análisis descriptivo de variables categóricas 19
Tabla 3: Análisis descriptivo de variables numéricas 22
Tabla 4: Resumen estadístico de la edad por grupos de depresión 24
Tabla 5: Relación entre actividad física y depresión 27
Tabla 6: Relación entre sexo y depresión 27
Tabla 7: Estadísticos del consumo medio de alcohol y del peso con y sin outliers 30
Tabla 8: Número y porcentaje de valores perdidos por variable 31
Tabla 9: Importancia de las variables explicativas de la severidad depresiva según información mutua 36
Tabla 10: Métricas de evaluación de los modelos 37
Tabla 11: Desempeño comparativo de los modelos de clasificación según las métricas de Exactitud, Precisión de la clase positiva, Sensibilidad, F1-score, AUC y la matriz de confusión expresada como [TN, FP; FN, TP] 39
Tabla 12: Métricas de evaluación de reglas de asociación 47
Tabla 13: Reglas de asociación más relevantes para la identificación de depresión severa 50

Índice de ilustraciones

Ilustración 1: Representación del modelo XGBoost 39

RESUMEN

Este trabajo tiene como objetivo analizar los factores asociados a la depresión en la población adulta española a partir de la Encuesta de Salud de España 2023, elaborada por el Instituto Nacional de Estadística, con una muestra de 21.032 individuos. La variable objetivo es la presencia o ausencia de depresión, construida a partir de la variable SEVERIDAD_DEPRESIVA y recodificada en formato dicotómico.

La metodología empleada combina análisis descriptivo, selección de variables mediante información mutua, aplicación de técnicas de aprendizaje automático, análisis de interpretabilidad de los resultados y reglas de asociación para identificar perfiles de mayor riesgo.

Los resultados muestran que las variables con mayor capacidad explicativa fueron la edad, la actividad física, la Comunidad Autónoma de residencia y el consumo de tranquilizantes. Desde una perspectiva preventiva, uno de los hallazgos más relevantes es el papel de la actividad física como principal elemento protector frente a la depresión, incluso por encima de variables biológicas o demográficas sobre las que existe menor margen de intervención. Además, el análisis de perfiles de riesgo permitió identificar un grupo especialmente vulnerable: mujeres mayores de 55 años, sedentarias y consumidoras de tranquilizantes, cuya prevalencia de depresión severa es casi cuatro veces superior a la observada en la población general.

En conjunto, los resultados confirman la naturaleza multifactorial de la depresión y apuntan a la necesidad de estrategias de prevención basadas en la promoción de la actividad física, el seguimiento de pacientes consumidores de tranquilizantes y un enfoque diferenciado por sexo en los protocolos de detección temprana.

PALABRAS CLAVE

Depresión, aprendizaje automático, información mutua, factores de riesgo, España

ABSTRACT

This study aims to analyse the factors associated with depression in the Spanish adult population using data from the 2023 Spanish Health Survey, conducted by the Spanish National Statistics Institute, with a sample of 21,032 individuals. The target variable is the presence or absence of depression, constructed from the variable SEVERIDAD_DEPRESIVA and recoded into a binary format.

The methodology combines descriptive analysis, feature selection through mutual information, the application of machine learning techniques, interpretability analysis of the results and association rules to identify higher-risk profiles.

The results show that the variables with the greatest explanatory capacity were age, physical activity, Autonomous Community of residence and the use of tranquillisers. From a preventive perspective, one of the most relevant findings is the role of physical activity as the main protective factor against depression, even above biological or demographic variables over which there is less scope for intervention. In addition, the risk profile analysis identified a particularly vulnerable group: women over the age of 55 who are sedentary and use tranquillisers, whose prevalence of severe depression is almost four times higher than that observed in the general population.

Overall, the findings confirm the multifactorial nature of depression and point to the need for prevention strategies based on the promotion of physical activity, the monitoring of patients who use tranquillisers and a sex-specific approach in early detection protocols.

KEY WORDS

Depression, machine learning, mutual information, risk factors, Spain

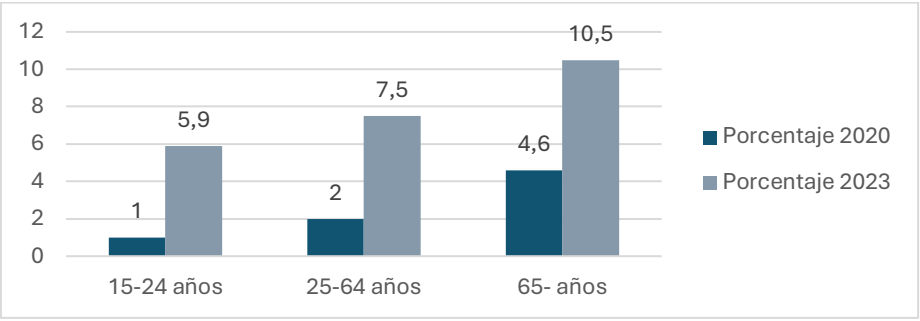
CAPÍTULO I: INTRODUCCIÓN

1. CONTEXTO Y RELEVANCIA DE ESTUDIO

La depresión es uno de los trastornos psiquiátricos más frecuentes y de mayor impacto en la población. Se caracteriza principalmente por un ánimo decaído, acompañado de una notable pérdida de interés o placer en las actividades cotidianas. Esta condición afecta tanto al ámbito físico, donde pueden producirse cambios en el peso o el apetito, como a la esfera psicológica, con síntomas como problemas de memoria, dificultades de concentración y tristeza persistente (Retamal, 1999, p. 9).

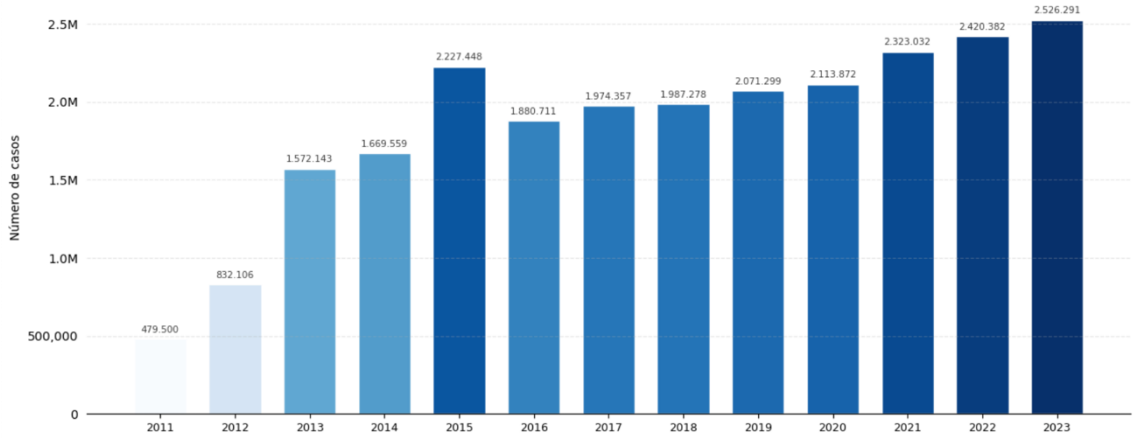
Según la OMS (2025), la depresión afecta a alrededor de **332 millones de personas**, lo que equivale a cerca del **4 % de la población mundial**. En el contexto español, la magnitud del problema es igualmente alarmante. De acuerdo con el Instituto Nacional de Estadística (INE, 2025), en los últimos años la prevalencia de episodios graves de depresión ha aumentado considerablemente. Como se muestra en el Gráfico 1, el porcentaje de personas mayores de 15 años con depresión grave aumentó entre 2020 y 2023 en todos los grupos de edad: del 1 % al 5,9 % en la población joven (15 a 24 años), del 2 % al 7,5 % en adultos (25 a 64 años) y del 4,6 % al 10,5 % en personas de 65 y más años. Este incremento refleja un deterioro significativo del bienestar psicológico de la población. Esta tendencia se confirma también en la evolución del número de casos registrados de depresión o trastornos depresivos en España, que alcanzó en 2023 su valor más elevado (Ver Gráfico 2).

Gráfico 1: Variación porcentual de cuadros depresivos severos activos por grupo etario en España (2020 y 2023)



Fuente: INE (2025)

Gráfico 2: Número de casos de depresión o trastornos depresivos registrados en España de 2011 a 2023



Fuente: IM Médico (2025)

Además de su elevada prevalencia y su impacto en la población, la depresión se caracteriza por su **naturaleza multifactorial**. Su aparición y desarrollo no pueden atribuirse a una única causa aislada, sino que responden a la interacción de factores ambientales, sociales, psicológicos y biológicos que, en conjunto, aumentan la probabilidad de padecerla (OMS, 2025). Entre ellos destacan los hábitos de vida, la edad, el sexo, las condiciones de trabajo, la salud y el entorno social. Esta complejidad pone de manifiesto la necesidad de abordar la depresión desde una perspectiva integral, capaz de considerar simultáneamente las distintas dimensiones que intervienen en su desarrollo. En este contexto, el uso de técnicas de aprendizaje automático (*machine learning*) resulta especialmente adecuado, ya que permite analizar de forma conjunta un gran número de variables y detectar patrones complejos que difícilmente pueden captarse mediante métodos estadísticos tradicionales.

En este contexto se sitúa el presente trabajo, que analiza la depresión en la población adulta española a partir de la Encuesta de Salud de España 2023 (ESdE 2023), una base de datos pública, amplia y representativa elaborada por el INE. La base utilizada contiene **21.032 observaciones y 432 variables**, lo que ofrece un marco adecuado para estudiar la depresión desde una perspectiva multifactorial. La variable objetivo del estudio es la depresión, entendida como la presencia o ausencia de sintomatología depresiva. Como variables explicativas se seleccionaron la edad, el sexo, la actividad física, el consumo de tranquilizantes, el apoyo social percibido, el consumo medio diario de alcohol, el hábito tabáquico, el peso, el estado civil, la convivencia en pareja, la Comunidad Autónoma de

residencia, la presencia de lesiones permanentes derivadas de accidentes, la existencia de tumores malignos, el nivel de estrés laboral y la jornada laboral.

Los resultados confirman que la depresión no puede explicarse a partir de un único factor, sino que es el resultado de la combinación de múltiples dimensiones. En la fase de selección de variables mediante información mutua, ninguna variable superó el umbral de capacidad explicativa establecido, lo que evidencia que ningún predictor aislado es suficiente para explicar el fenómeno. No obstante, las variables con **mayor capacidad explicativa** fueron la **edad**, la **actividad física**, la **Comunidad Autónoma** de residencia y el **consumo de tranquilizantes**. Además, el análisis descriptivo evidenció una mayor prevalencia de depresión en personas de mayor edad, en mujeres, en individuos sedentarios y en determinadas Comunidades Autónomas, lo que pone de manifiesto una clara heterogeneidad geográfica en su distribución.

La aplicación de modelos de aprendizaje automático permitió profundizar en este análisis, identificando patrones más complejos en los datos. Entre los modelos estimados, **XGBoost** fue el que mostró el mejor equilibrio en la detección de casos de depresión. A su vez, el análisis de interpretabilidad mediante SHAP reveló que la actividad física es el factor con mayor influencia en la predicción, seguida por la edad y el consumo de tranquilizantes. En cuanto a la dirección de sus efectos, niveles bajos de actividad física, una mayor edad y el consumo de tranquilizantes aumentan la probabilidad de depresión, mientras que una actividad física elevada actúa como factor protector. Por último, el análisis de reglas de asociación permitió identificar perfiles de riesgo concretos: el **perfil de máximo riesgo, mujer mayor de 55 años, sedentaria y consumidora de tranquilizantes**, presenta una prevalencia de depresión severa casi cuatro veces superior a la de la población general.

En definitiva, este estudio aporta una visión más completa del fenómeno al combinar una base de datos poblacional amplia con técnicas que no solo permiten predecir, sino también interpretar el papel de los distintos factores, contribuyendo así a una mejor comprensión de la naturaleza multifactorial de la depresión en España.

2. ANTECEDENTES

En la actualidad, el uso de **métodos de *Machine Learning*** en el ámbito médico ha crecido notablemente, especialmente en el estudio de los trastornos psicológicos debido a su naturaleza compleja y multifactorial. Estas técnicas permiten **procesar y examinar datos clínicos masivos** con el fin de detectar patrones relevantes en la salud mental de los pacientes, contribuyendo así a mejorar su diagnóstico y facilitar su tratamiento (Hasan, Shifat, Matubber et al., 2026).

Entre los modelos más utilizados en la literatura destacan los basados en **árboles de decisión**, las **redes neuronales** y las **técnicas de clustering** (Shatte, Hutchinson y Teague, 2019). En esta línea, Sánchez-Carro, de la Torre-Luque, Leal-Leturia, et al. (2022) emplearon un algoritmo de máquinas de vectores de soporte para clasificar a pacientes con trastorno depresivo mayor atendidos en centros sanitarios de Madrid, entre ellos el Hospital Universitario La Princesa, a partir de biomarcadores inmunometabólicos, marcadores de estrés oxidativo y hábitos de vida. Sus resultados identificaron la actividad física, el consumo de alcohol y el tabaquismo como algunas de las variables con mayor capacidad discriminativa según el área bajo la curva ROC. Asimismo, Lu, Wan y Liu (2025) analizaron los determinantes de los síntomas depresivos en adultos de mediana edad y mayores a partir de cinco cohortes internacionales, observando un mejor rendimiento de *XGBoost* y *LightGBM*. En el contexto de la pandemia, Simjanoski, Ballester, da Mota, et al. (2022) desarrollaron modelos de machine learning entrenados sobre una muestra española para predecir depresión y ansiedad a partir de comportamientos relacionados con el estilo de vida, como la actividad física, el consumo de tabaco y el consumo de alcohol. Para ello, emplearon *Elastic Net*, *Random Forest* y *Gradient Tree Boosting*.

No obstante, pese al creciente uso de técnicas de aprendizaje automático en el estudio de la depresión, todavía son escasos los trabajos que incorporan **metodologías avanzadas para la selección de variables**. En este contexto, el presente trabajo introduce el uso de la **información mutua (*mutual information*)** como técnica de selección de características, cuya implementación y justificación teórica se desarrollarán en detalle en el capítulo correspondiente.

Otra limitación habitual en la literatura es el uso de **muestras muy específicas**, como estudiantes universitarios, lo que **dificulta la generalización de los resultados**. La homogeneidad de estas poblaciones reduce la validez externa de los modelos, ya que los patrones identificados pueden no ser representativos de otros grupos de edad, contextos sociales o perfiles clínicos. Ejemplos de ello son los estudios de Wei, Qin, Liu, et al. (2025) y Baba y Bunji (2023), centrados en la predicción de problemas de salud mental en estudiantes universitarios de China y Japón, respectivamente.

En el entorno clínico, la **interpretabilidad** resulta esencial, ya que los modelos deben proporcionar explicaciones claras y fiables para apoyar la toma de decisiones. En este contexto, **SHAP** se ha consolidado como una de las herramientas más empleadas, al permitir comprender el efecto de cada variable en las predicciones (Trelles, Fontaines Ruiz y Ponce Rojo, 2025). En esta línea, Al Masud, Shanto, Sakin y Kabir (2025) aplicaron SHAP en un estudio sobre detección de depresión en estudiantes universitarios de Bangladesh. Tras comparar varios modelos, Random Forest obtuvo el mejor rendimiento, y SHAP permitió identificar la ansiedad y la soledad como los factores con mayor impacto en la predicción.

Asimismo, las **técnicas de minería de datos** basadas en **reglas de asociación** han sido aplicadas en el ámbito de la salud mental para **identificar perfiles de riesgo** y **combinaciones de factores** asociados a trastornos como la depresión (Biilah et al., 2021; Wang et al., 2023).

3. OBJETIVOS

El objetivo principal de esta investigación es **analizar los factores de riesgo más relevantes asociados a la depresión en la población adulta española**, utilizando datos de la **Encuesta de Salud de España 2023**. Para ello, se adopta un enfoque multifactorial que integra variables sociodemográficas, clínicas y relacionadas con los hábitos de vida.

La **variable objetivo del estudio es la depresión**, construida a partir de la variable SEVERIDAD_DEPRESIVA de la ESdE 2023. Se trata de una **variable derivada** que cuantifica la severidad de la sintomatología depresiva a partir de las respuestas del propio encuestado. Al basarse en autoinforme y no en un diagnóstico clínico emitido por un

profesional sanitario, la variable refleja la sintomatología percibida por el informante, sesgo que se aborda en el apartado de limitaciones. Dado que esta variable presenta originalmente cinco niveles de severidad y una elevada concentración de casos en la categoría de ausencia de depresión, se ha **recodificado en formato dicotómico**, diferenciando entre presencia y ausencia de depresión.

Se han empleado las siguientes **variables explicativas**: el consumo medio diario de alcohol, el nivel de apoyo social percibido, el hábito tabáquico, la actividad física, el peso, el consumo de tranquilizantes, el nivel de estrés laboral, la jornada laboral, el estado civil, la convivencia en pareja, el sexo, la edad, la comunidad autónoma de residencia, la presencia de lesiones permanentes derivadas de accidentes y la existencia de tumores malignos.

Los objetivos específicos son los siguientes:

- Conocer la **relación entre la depresión y factores sociodemográficos** como el sexo, la edad, el estado civil, la convivencia en pareja y el contexto geográfico de residencia.
- Aplicar la técnica de **selección de variables *Mutual Information*** para identificar qué variables presentan mayor capacidad explicativa respecto a la presencia de depresión y guiar así, de forma fundamentada, el proceso de modelización.
- **Explicar la variable objetivo** a partir del menor número posible de variables asociadas, determinando cuáles y en qué dirección contribuyen a la probabilidad de padecer depresión.
- Descubrir qué **combinaciones de factores** configuran perfiles de mayor riesgo de **depresión severa**.
- **Contrastar empíricamente** algunas **hipótesis presentes en la literatura**, como la mayor prevalencia de depresión en mujeres o la influencia de la edad y el estilo de vida.
- **Contribuir al estado actual de la cuestión** proporcionando una mejor comprensión de la naturaleza multifactorial de la depresión en España.

4. METODOLOGÍA

La metodología empleada en este trabajo se estructuró en diferentes fases que garantizan el rigor y la coherencia del análisis.

Para entender cuál es la situación actual, primero se realizó un **análisis detallado de bibliografía** utilizando fuentes académicas de Google Académico. Esta etapa contribuyó a determinar las herramientas metodológicas más significativas y los autores de referencia en la investigación acerca de la depresión. Este estudio bibliográfico evidenció que la depresión continúa siendo un tema de gran importancia desde el punto de vista social y científico, y también señaló las **limitaciones de los modelos estadísticos tradicionales** para captar la complejidad multifactorial de esta condición mental. Esto último explica y motiva el uso de **técnicas de aprendizaje automático**.

Posteriormente, se seleccionó una fuente de datos pública, fiable y representativa como es la **Encuesta de Salud de España (ESdE)** elaborada por el **Instituto Nacional de Estadística (INE)**. Esta base de datos es apropiada para examinar la depresión desde un enfoque multifactorial, pues incluye todo tipo de variables sobre el estado de salud, los hábitos de vida y los rasgos sociodemográficos de los españoles. Con base en esta fuente, se llevó a cabo un procedimiento de **limpieza y preprocesamiento** de la información que abarcó la redefinición de respuestas, la identificación y tratamiento de valores extremos y la imputación de los valores ausentes.

Para las fases posteriores de **análisis y modelización** se utilizó el lenguaje de programación **Python**, apoyándose en diversas **librerías** especializadas. El tratamiento y manipulación de los datos se realizó mediante ***pandas*** (The pandas development team, 2020) y ***NumPy*** (Harris et al., 2020). En primer lugar, se aplicó la técnica de Información Mutua, implementada a través de ***scikit-learn*** (Pedregosa et al., 2011), con el fin de seleccionar las variables con mayor capacidad explicativa respecto a la presencia de depresión. Posteriormente, se implementaron y evaluaron dos modelos de clasificación: la **regresión logística**, incluida también en *scikit-learn* y ***XGBoost*** (Chen y Guestrin, 2016). El rendimiento de ambos modelos se evaluó mediante diversas métricas de desempeño y se comparó con un **modelo trivial** como referencia, con el objetivo de identificar el modelo con mayor capacidad explicativa del fenómeno depresivo. Una vez

seleccionado dicho modelo, se incorporó un **análisis de interpretabilidad mediante SHAP** (*SHapley Additive exPlanations*) (Lundberg y Lee, 2017), con el fin de determinar en qué dirección y con qué intensidad contribuye cada variable a la predicción. Adicionalmente, se aplicó el **algoritmo Apriori** a través de ***mlxtend*** (Raschka, 2018) para la obtención de **reglas de asociación**. La **visualización** de resultados se llevó a cabo con ***matplotlib*** (Hunter, 2007) y ***seaborn*** (Waskom, 2021).

Como parte de la metodología se han mantenido reuniones periódicas y sesiones de Python con el director del TFG.

5. ESTRUCTURA

Para cumplir con los objetivos establecidos en la sección previa y tratar el análisis multifactorial de la depresión, este trabajo está organizado en **siete capítulos.**

El primer capítulo se enfoca en la introducción del tema que se estudia, subrayando lo relevante y pertinente que es examinar la depresión desde un **punto de vista multifactorial**. Este capítulo también muestra los antecedentes, los objetivos y el método de la investigación. En el segundo capítulo se aborda el **tratamiento y preprocesamiento de datos**, incluyendo la **selección de las variables** más pertinentes para el estudio de la depresión a partir de la literatura científica, además de un **análisis descriptivo** preliminar que facilita una comprensión más profunda de estas variables. El tercer capítulo se centra en el uso de una **técnica sofisticada para seleccionar variables, llamada información mutua** (*mutual information*), con el propósito de determinar qué variables impactan más en la presencia de la depresión en los adultos españoles. El cuarto capítulo se enfoca en el diseño de diversos **modelos de aprendizaje automático en Python**. También se evalúa su **capacidad predictiva** para determinar cuál es el modelo más eficaz. El quinto capítulo se enfoca en la **técnica SHAP para interpretar** el modelo con mejor desempeño. El sexto capítulo aplica el **algoritmo Apriori** para la obtención de **reglas de asociación**, con el objetivo de identificar combinaciones de factores que configuran perfiles de mayor riesgo de depresión severa. El último capítulo incluye las **conclusiones** clave derivadas del análisis y sugiere recomendaciones e investigaciones futuras.

CAPÍTULO II: TRATAMIENTO DE LOS DATOS

1. BASE DE DATOS UTILIZADA

Para llevar a cabo la presente investigación, se ha decidido emplear una **fuente de datos pública, confiable y anonimizada**: la **Encuesta de Salud de España (ESdE) 2023, elaborada por el INE**. Esta encuesta está enfocada en los ciudadanos que viven en viviendas familiares de todo el país y obtiene datos específicos acerca del bienestar físico, los hábitos de vida, las condiciones de trabajo y múltiples aspectos sociodemográficos (Ministerio de Sanidad, 2023).

La base de datos fue consultada y descargada directamente desde la página oficial del INE, donde se encuentra disponible en formato CSV (*Comma-Separated Values*) para su tratamiento y análisis estadístico. La ESdE 2023 es una encuesta de ámbito nacional con una muestra de aproximadamente 37.500 viviendas distribuidas en 2.500 secciones censales, diseñada para ser representativa de la población adulta española tanto a nivel nacional como por Comunidad Autónoma. En cada vivienda se seleccionó aleatoriamente **un adulto de 15 o más años** para cumplimentar el cuestionario, obteniéndose un total de **21.032 respuestas válidas**.

2. PRESELECCIÓN DE VARIABLES

La base de datos original contenía **432 variables**, lo que dificulta la creación de modelos eficaces y precisos. Asimismo, dado que se trata de una encuesta general de salud, numerosas variables recopiladas no presentan una relación directa con la salud mental ni, en particular, con la depresión. Por esta razón, se llevó a cabo un **proceso de preselección de variables basado en criterios teóricos y empíricos**. En concreto, se seleccionaron aquellas variables que, según la literatura científica previa, han mostrado una asociación relevante con la depresión, así como aquellas que permiten analizar el fenómeno desde un enfoque multifactorial.

En particular, se incorporaron variables de tres tipos: **sociodemográficas** (como la edad, el sexo o la comunidad autónoma), variables relacionadas con los **estilos de vida** (como el consumo de alcohol, el tabaquismo o la actividad física) y variables **clínicas** (como el consumo de tranquilizantes o la presencia de limitaciones físicas o enfermedades). Esta selección responde a la necesidad de captar la complejidad de la depresión, entendida como un fenómeno influido por la interacción de múltiples factores. Este proceso

permitió **reducir la dimensionalidad** del conjunto de datos y trabajar con un conjunto de variables coherente con los objetivos del estudio, centrado en identificar y analizar los principales factores asociados a la depresión en la población adulta española.

Las variables seleccionadas se renombraron mediante Python para facilitar la comprensión e interpretación de los resultados en las fases posteriores del análisis.

A continuación, se presentan un total de 16 variables preseleccionadas, indicando su nombre original en la base de datos, el nombre modificado, su descripción y su relevancia teórica o empírica en el estudio de la depresión.

Tabla 1: Variables preseleccionadas y su relevancia teórica

Variable original	Nombre modificado	Descripción	Relevancia en el estudio
SEVERIDAD DEPRESIVA	**Depresión**	Ausencia o presencia de depresión	Variable dependiente/objetivo del estudio. Permite clasificar los individuos con y sin depresión.
CMD1	**consumo_medio_diario_alcohol**	Consumo medio de alcohol semanal (gramos)	El consumo de alcohol ha mostrado asociación con mayores tasas de depresión (Rosales-Damián et al., 2024).
S1	**apoyo_social_percibido**	Nivel de apoyo social percibido por el individuo	El bajo apoyo social se asocia consistentemente con síntomas depresivos y mayor vulnerabilidad psicológica (Jiménez-Hernández et al., 2022).
Q1	**fumador**	Indica si la persona fuma actualmente	Diversos estudios relacionan el tabaquismo con mayor prevalencia de depresión (Fluharty et al., 2017).
O2	**actividad_física**	Frecuencia de actividad física moderada o intensa	La inactividad física está relacionada con mayor riesgo de depresión, mientras que el ejercicio regular actúa como factor protector (Choi et al., 2019).
N2	**peso**	Peso corporal en kilogramos	El peso y el IMC se han vinculado con la depresión (Alonso y Olivos, 2020).
Uk3_7a	**consumo_tranquilizantes**	Uso de tranquilizantes, relajantes o pastillas para dormir	El uso de psicofármacos y relajantes puede reflejar un cuadro previo o coexistente de depresión por ser medicamentos que se recetan para disminuir los efectos de la misma.
H3	**nivel_estres_laboral**	Nivel de estrés laboral percibido (1=nada, 7=muy estresante)	Según Celso Arango, jefe del Servicio de Psiquiatría del Hospital Gregorio Marañón de Madrid, "el estrés aumenta el cortisol, que es neurotóxico. El estrés crónico, mantenido, acaba produciendo insomnio, ansiedad y cuadros depresivos". Además, advierte que, si se lograra reducir el nivel de estrés laboral, considerado un importante factor de riesgo, podrían disminuirse hasta un 18% los casos de depresión (Arango, citado en Mouzo, 2022).
B11	**jornada_laboral**	Tipo de jornada laboral (completa, parcial)	A medida que aumentan las horas de trabajo diario, también lo hace la probabilidad de experimentar síntomas de ansiedad y depresión, sobre todo, a partir de jornadas con más de 5 horas diarias (Lang et al., 2020).
A5	**estado_civil**	Estado civil (casado, soltero, viudo, separado)	Las personas que no tienen una pareja estable presentan una probabilidad un 80 % mayor de padecer depresión en comparación con quienes sí mantienen una relación estable (Mediavilla, 2024).
A4	**convivencia_en_pareja**	Si convive o no en pareja	Un reciente estudio reveló que las personas solteras presentan un 80% más de probabilidades de experimentar depresión en comparación con aquellas que tienen pareja (Zhai et al., 2024).
SEXOa	**sexo**	Sexo del individuo	Las mujeres presentan una mayor prevalencia de depresión que los hombres: un 6,9% frente a un 4,6% (OMS, 2025).

EDADa	**edad**	Edad del individuo	La edad es una variable clave, ya que la depresión afecta de forma distinta según la etapa vital, reflejando diferentes factores de riesgo y necesidades de prevención.
CCAA	**CCAA**	Comunidad autónoma de residencia	Permite captar posibles diferencias geográficas o contextuales en la prevalencia de depresión.
C5a_33	**lesiones_permanentes_accidente**	Lesiones permanentes derivadas de accidentes	Las limitaciones físicas derivadas de lesiones pueden aumentar el riesgo de depresión (O'Hagan et al., 2013).
C5a_27	**tumor_maligno**	Diagnóstico de tumores malignos	Se estima que entre el 15 % y el 25 % de los pacientes oncológicos presentan algún grado de depresión. (Rodríguez et al., 2015).

Fuente: Elaboración propia

3. ANÁLISIS DESCRIPTIVO

Para asegurar la validez y coherencia de los resultados, fue preciso llevar a cabo una depuración preliminar de los datos antes de comenzar el análisis descriptivo. Se detectaron en la base de datos ciertas codificaciones que no reflejaban valores reales, sino la falta de respuesta de los encuestados. Específicamente, se consideraron las categorías que tenían el valor **9**, que representan **"no contesta",** y también el valor **999** en las variables de peso y consumo medio diario de alcohol como información no disponible. Posteriormente, fueron **recodificadas como valores perdidos (NaN)** para evitar distorsiones en los análisis estadísticos posteriores.

Después de haber llevado a cabo esta depuración inicial, se realizó un análisis descriptivo de las variables elegidas para entender la estructura y los rasgos más importantes del conjunto de datos previo a la etapa de modelización. Esta evaluación posibilitó tener una perspectiva general de cómo se distribuyen las variables, reconocer desequilibrios entre las categorías y examinar la presencia de valores perdidos en cada variable para poder tratarlos en el siguiente punto del presente capítulo. Asimismo, con el fin de profundizar en el conocimiento de los datos, se analizará la relación entre la variable objetivo (depresión) y determinadas variables explicativas relevantes, como la edad, la actividad física y la Comunidad Autónoma, mediante el uso de tablas y representaciones gráficas. Este análisis permitirá identificar posibles patrones y asociaciones preliminares que servirán de base para el posterior desarrollo de los modelos predictivos.

3.1 Análisis de las variables categóricas

Para las variables categóricas, se calcularon tanto la frecuencia absoluta como la relativa de cada categoría, además del total y el porcentaje de valores ausentes que estaban relacionados (Ver tabla 2).

Tabla 2: Resumen del análisis descriptivo de variables categóricas

Variable	Categorías	Frecuencia	Proporción (%)	Valores perdidos	Valores perdidos (%)
Depresión	Ninguna Leve Moderada Moderadamente grave Grave	13915 3508 1648 879 383	66,16 16,68 7,84 4,18 1,82	699	3,32
Apoyo_social_percibido	Más de cinco personas Tres a cinco personas Una a dos personas Ninguna persona	8025 7723 4752 245	38,16 36,72 22,59 1,16	287	1,36
Fumador	No, nunca No actualmente, antes sí Si, a diario Si, no a diario	11967 4891 3564 497	56,90 23,26 16,95 2,36	113	0,54
Actividad_fisica	Actividad ocasional Sedentario Varias veces a la semana Varias veces al mes	8541 6574 3193 2453	40,61 31,26 15,18 11,66	271	1,29
Consumo_tranquilizantes	No Si	9920 3411	47,17 16,22	7701	36,62
Nivel_estres_laboral	Bastante estresante Muy estresante Moderadamente estresante Extremadamente estresante Algo estresante Poco estresante Nada estresante	2386 1944 1485 1306 1107 915 591	11,34 9,24 7,06 6,21 5,26 4,35 2,81	11298	53,72
Jornada_laboral	Tiempo completo Tiempo parcial	8798 986	41,83 4,69	11248	53,48
Estado_civil	Casado/a Soltero/a Viudo/a Divorciado/a Separado/a legalmente	9917 6321 2674 1433 552	47,15 30,05 12,71 6,81 2,62	135	0,64
Convivencia_en_pareja	No conviviendo en pareja Conviviendo con cónyuge Conviviendo con pareja de hecho	11009 9034 594	52,34 42,95 2,82	395	1,88
Sexo	Mujer Hombre	11352 9680	53,97 46,03	0	0
CCAA	Andalucía Comunidad Valenciana Madrid Cataluña País Vasco Castilla y León Extremadura Murcia Aragón Castilla-La-Mancha Galicia Canarias Navarra Asturias Islas Baleares Cantabria La Rioja Melilla	2674 2059 1876 1802 1364 1344 1099 1080 1044 1030 904 874 854 705 680 635 579 226	12,71 9,79 8,92 8,57 6,49 6,39 5,23 5,14 4,96 4,90 4,30 4,16 4,06 3,35 3,23 3,02 2,75 1,07	1802	8,57

	Ceuta	203	0,97		
Lesiones_permanentes_accidente	No Si	19726 1215	93,79 5,78	91	0,43
Tumor_maligno	No Si	19742 1100	93,87 5,23	190	0,90

Fuente: Elaboración propia

En la muestra analizada, las **mujeres representan el 53,97 % y los hombres el 46,03 %.** Más del **70 %** de las personas encuestadas afirma tener al menos tres figuras de referencia, por lo que la mayoría de los individuos **sostiene tener redes de apoyo social** extensas o moderadas. En cuanto al consumo de tabaco, **la mayoría de los encuestados nunca han fumado** y alrededor del 23% son exfumadores, en cambio, los fumadores actuales constituyen una porción más pequeña.

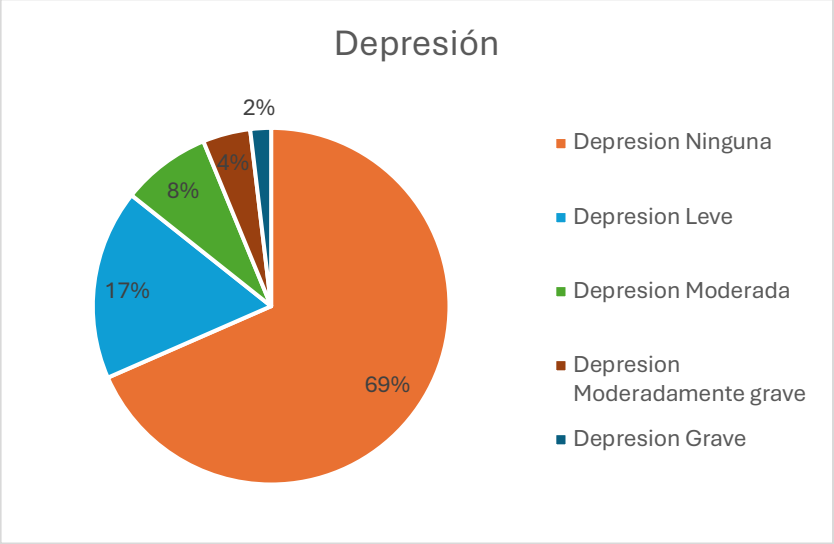
La actividad física presenta una distribución heterogénea, con más frecuencia la **actividad ocasional** y, después, un porcentaje significativo de individuos con un estilo de vida sedentario. La **mayoría de las personas dice no consumir tranquilizantes**, sin embargo, esta variable, como la jornada laboral, tiene un alto porcentaje de valores perdidos. Esto sugiere precaución al interpretarla y justifica un tratamiento particular en el siguiente apartado.

La distribución por Comunidades Autónomas concuerda con el peso poblacional, siendo **Andalucía, la Comunidad Valenciana, Madrid y Cataluña las que más destacan**. Por último, las variables clínicas específicas, **lesiones permanentes y tumores malignos** presentan una baja prevalencia, siendo **mayoritaria la ausencia** de estas condiciones en la muestra.

Con el propósito de estudiar la variable objetivo depresión más a fondo, se creó un gráfico circular (Ver Gráfico 3) para analizar cómo están distribuidas las respuestas y determinar

si sería conveniente recodificarlas. La variable depresión se presenta originalmente como una **escala ordinal de cinco niveles**, que va desde ausencia de depresión hasta depresión grave.

Gráfico 3: Distribución inicial de la variable objetivo depresión antes de la recodificación

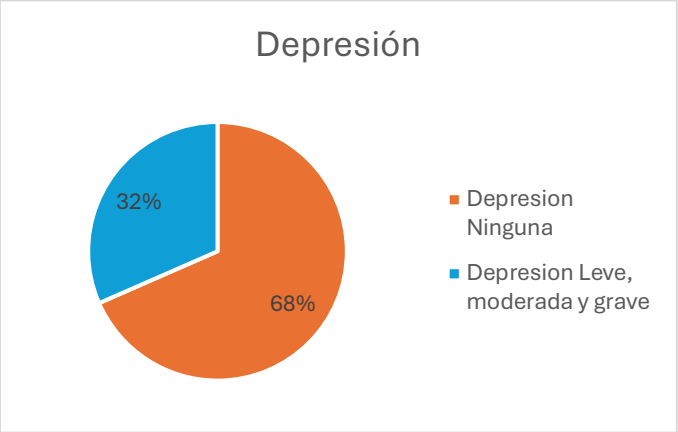


Fuente: Elaboración propia

El gráfico muestra una clara concentración de casos en la categoría de **ausencia de depresión**, que representa **aproximadamente el 70 %.** Esta marcada asimetría entre categorías justifica la **recodificación de la variable en formato dicotómico**, diferenciando entre ausencia de depresión y presencia de sintomatología depresiva (categorías leve a grave). Esta modificación posibilita **disminuir la desproporción** entre clases, evitar sesgos en favor de la categoría predominante y promover una mayor estabilidad en el desempeño predictivo de los modelos. Además, la predicción de niveles específicos de severidad (por ejemplo, distinguir entre depresión leve y moderadamente leve) resulta especialmente compleja y propensa a error, mientras que el enfoque binario facilita una identificación más robusta y fiable.

El Gráfico 4 muestra la distribución después de la recodificación y revela que la variable depresión sigue desequilibrada, con un 68 % de individuos sin depresión en comparación con el 32 % que sí la presentan. Al tratarse de un desbalanceo moderado, no se considera necesario el uso de técnicas específicas para datos desbalanceados, por lo que en el Capítulo IV se emplean modelos estándar comparados con un modelo trivial como referencia.

Gráfico 4: Distribución de la variable objetivo depresión tras la recodificación



Fuente: Elaboración propia

Tras el análisis descriptivo, se procedió a la recodificación de algunas variables con el objetivo de simplificar su estructura, reducir categorías con baja frecuencia y mejorar su adecuación para el análisis predictivo.

Las categorías originales de la variable "fumador" se reagruparon con el fin de facilitar su interpretación. Las respuestas “sí, a diario” y “sí, no a diario” se agruparon en una única categoría que indica consumo actual de tabaco, mientras que se mantuvieron diferenciadas las categorías “no actualmente, antes sí” (exfumador) y “no, nunca” (nunca fumador). Esta recodificación posibilita disminuir la fragmentación de la variable y examinar con mayor claridad la relación entre el consumo de tabaco y la existencia de depresión.

3.2 Análisis de las variables numéricas

En cuanto a las variables numéricas (edad, peso y consumo medio de alcohol al día), se calcularon las principales medidas descriptivas, como la **media**, la **desviación estándar** y los **percentiles 25 y 75,** con el fin de comprender su distribución.

Tabla 3: Análisis descriptivo de variables numéricas

Variables	Media	Desviación estándar	Primer cuartil (25%)	Tercer cuartil (75%)	Valores perdidos	Valores perdidos (%)
Edad (años)	54,49	19,14	40	69	0	0
Peso (Kg)	72,97	14,33	63	81	867	4,12

Consumo_medio_ diario_alcohol (g)	3,95	8,87	0	4,29	51	0,24

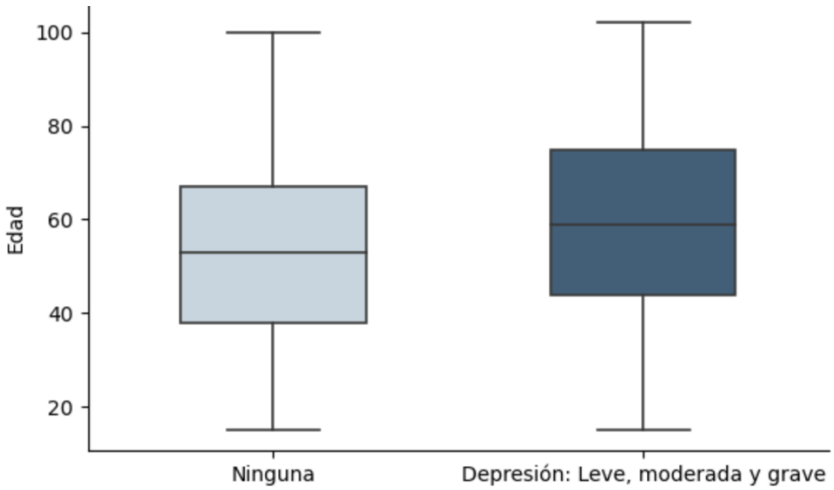
Fuente: Elaboración propia

La Tabla 3 muestra que la edad media de la muestra es de 54,49 años. El peso presenta una media de 72,97 kg y una desviación estándar relativamente elevada, lo que indica una notable variabilidad entre los individuos. Por su parte, el consumo medio diario de alcohol es bajo (3,95 g), aunque acompañado de una desviación estándar alta, lo que sugiere una distribución asimétrica caracterizada por una mayoría de individuos con consumo nulo o muy reducido y un grupo minoritario con consumos más elevados. El análisis detallado de la distribución y de los posibles valores atípicos de estas variables se desarrollará en el capítulo siguiente.

3.3 Relación entre la depresión y variables explicativas

En primer lugar, se analizó la relación entre la depresión y la edad. Como se observa en el Gráfico 5 y la Tabla 4, los individuos que presentan depresión tienden a tener una edad ligeramente superior en comparación con aquellos que no presentan esta condición. La edad media de los individuos con depresión es de 58,62 años, frente a los 52,40 años en aquellos sin depresión. Asimismo, la mediana también es mayor en el grupo con depresión (59 años frente a 53 años), lo que sugiere una **mayor prevalencia de sintomatología depresiva en edades más avanzadas**.

Gráfico 5: Edad según presencia de depresión



Fuente: Elaboración propia

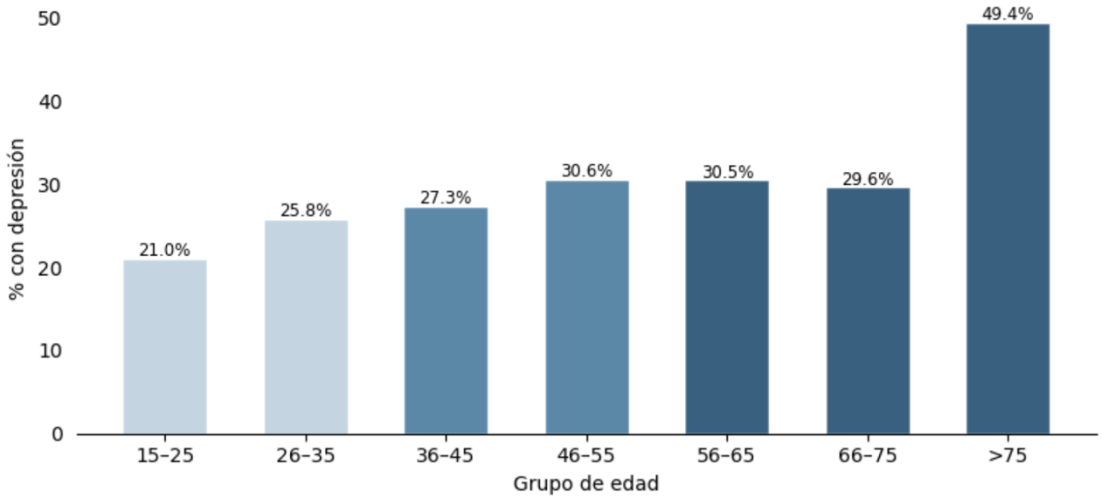
Tabla 4: Resumen estadístico de la edad por grupos de depresión

Depresión	Media	Mediana	Mínimo	Máximo
Depresión: leve, moderada y grave	58,62	59	15	102
Ninguna	52,40	53	15	100

Fuente: Elaboración propia

El Gráfico 6 desglosa la prevalencia por grupos de edad, confirmando la tendencia ascendente: desde el 21,0 % en los jóvenes de 15 a 25 años hasta el 49,4 % en mayores de 75. Cabe destacar que la prevalencia se estabiliza en torno al 30 % entre los 46 y los 75 años, para experimentar un incremento pronunciado a partir de dicha franja.

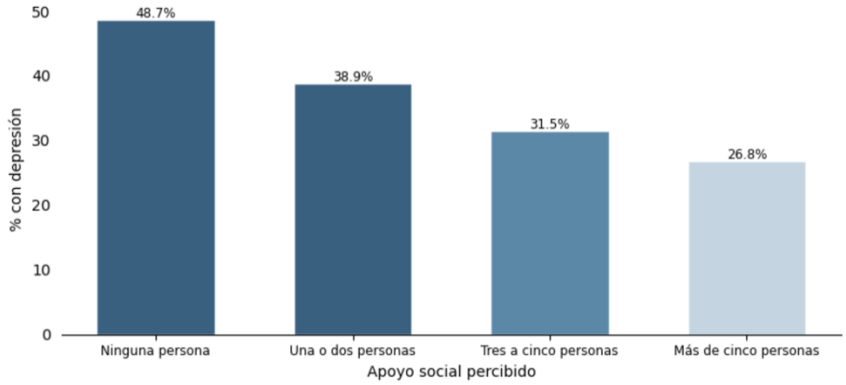
Gráfico 6: Porcentaje de depresión por grupo de edad



Fuente: Elaboración propia

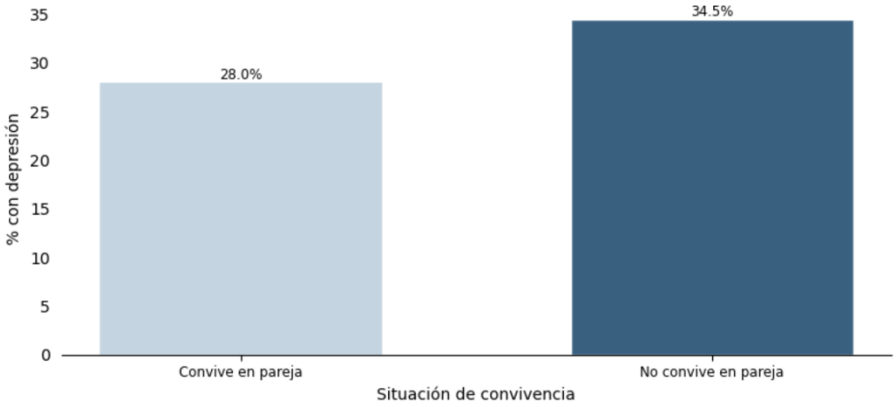
Junto a la edad, resultó de interés analizar en qué medida el entorno afectivo y social del individuo se relaciona con la presencia de depresión, dado que la calidad de los vínculos cercanos puede condicionar tanto la aparición como la evolución de los síntomas depresivos.

Gráfico 7: Prevalencia de depresión por apoyo social percibido



Fuente: Elaboración propia

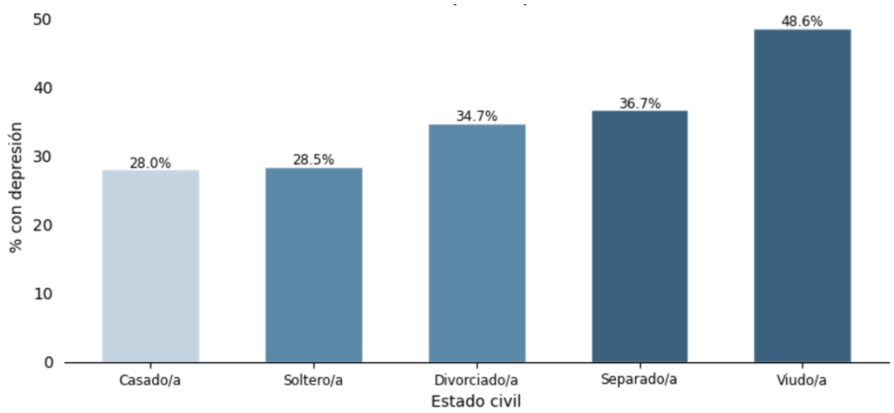
Gráfico 8: Prevalencia de depresión por convivencia en pareja



Fuente: Elaboración propia

Los Gráficos 7 y 8 muestran una tendencia común: tanto el **apoyo social percibido como la convivencia en pareja se asocian con una menor prevalencia de depresión**. Quienes no cuentan con ninguna persona de apoyo presentan una tasa del 48,7 %, frente al 26,8 % de quienes disponen de más de cinco personas; del mismo modo, no convivir en pareja se asocia a una prevalencia del 34,5 %, frente al 28,0 % de quienes sí conviven.

Gráfico 9: Prevalencia de depresión por estado civil



Fuente: Elaboración propia

En línea con lo anterior, el Gráfico 9 muestra que **los procesos de pérdida de pareja** (ya sea por viudedad (48,6 %), separación (36,7 %) o divorcio (34,7 %)) se asocian con las **mayores tasas de depresión**. Por el contrario, las personas casadas y solteras presentan prevalencias similares y más bajas, en torno al 28 %.

Con el propósito de medir la relación entre la variable objetivo (depresión) y distintas variables categóricas binarias, se crearon **tablas de contingencia y se determinó el *odds ratio* (OR).** Esta medida permite comparar la probabilidad de que ocurra un evento en dos grupos. Para calcularlo se utiliza una tabla de contingencia, en la que se separan a las personas expuestas y no expuestas a un factor específico, así como la presencia o ausencia del evento de interés, en este caso, la depresión. El OR se define matemáticamente como el **cociente de los productos cruzados de la tabla mencionada**. Un valor de OR igual a 1 significa que no hay una relación entre las variables. Un **valor por encima de 1 señala que el evento es más probable en el grupo expuesto**, en cambio un valor por debajo de 1 indica que es menos probable, lo cual podría entenderse como un efecto protector (BioDatev, 2024).

La Tabla 5 muestra la relación entre la actividad física y la presencia de depresión. Se observa que los **individuos sedentarios presentan una mayor proporción de depresión** en comparación con aquellos que realizan actividad física. En concreto, entre los individuos sedentarios, 2605 presentan depresión frente a 3742 que no la presentan, mientras que en el grupo que realiza actividad física, 3743 individuos presentan depresión frente a 10072 que no la presentan. Para cuantificar esta asociación, se calculó el *odds*

ratio, obteniéndose un valor de OR = 0,534. Dado que este valor es inferior a 1, se concluye que la **actividad física se asocia con una menor probabilidad de presentar depresión.** En términos relativos, los individuos que realizan actividad física tienen aproximadamente un 46,6 % (1-0,534) menos de probabilidad de sufrir depresión que aquellos sedentarios, lo que sugiere un efecto protector de este factor.

Tabla 5: Relación entre actividad física y depresión

	No depresión	Depresión
Sedentario	3742	2605
Activo	10072	3743

Fuente: Elaboración propia

Asimismo, se analizó la relación entre el sexo y la presencia de depresión. La Tabla 6 muestra que la proporción de individuos con depresión es mayor en mujeres (37,14 %) que en hombres (25,05 %). Se obtuvo un valor de OR = 0,566. Este resultado indica que los **hombres presentan una menor probabilidad de sufrir depresión** en comparación con las mujeres, en coherencia con los porcentajes observados. Este resultado se encuentra en línea con estudios previos que señalan que las mujeres presentan una mayor predisposición a padecer trastornos depresivos (Montesó-Curto y Aguilar-Martín, 2013; OMS, 2025). La interacción de diversos factores sociales, psicológicos y biológicos puede ser la causa de que las mujeres tengan una prevalencia más alta de depresión. Desde el punto de vista biológico, las mujeres experimentan cambios hormonales a lo largo de todas las etapas vitales, como la pubertad, el embarazo o la menopausia. Estos cambios afectan a su estado de ánimo y las hacen más propensas a sufrir depresión. Además, los factores de tipo psicosocial juegan un rol esencial. Las mujeres tienen más probabilidad de enfrentar situaciones de estrés crónico, como la violencia de género, la desigualdad económica y social o el exceso de responsabilidades familiares en el hogar. Esto aumenta las posibilidades de que sufran depresión (Mayo Clinic, 2025).

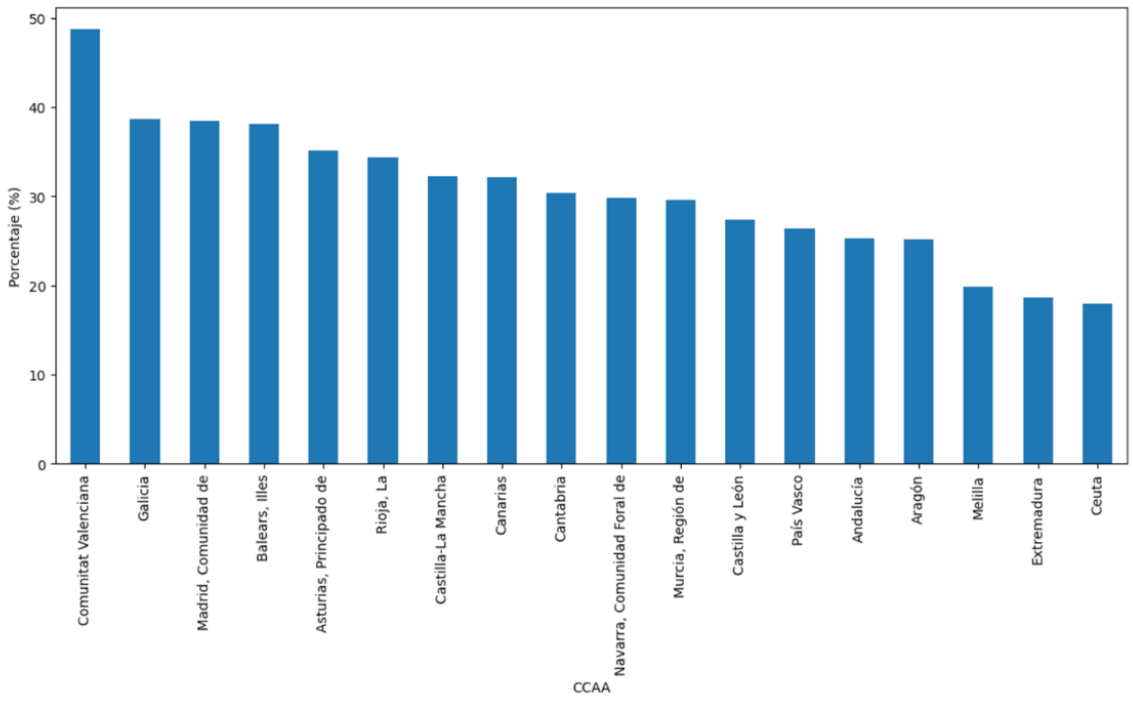
Tabla 6: Relación entre sexo y depresión

	No depresión	Depresión
Mujer	6885	4068
Hombre	7030	2350

Fuente: Elaboración propia

Por último, se analizó la distribución de la depresión según la Comunidad Autónoma. Los resultados muestran diferencias notables entre regiones (Ver Gráfico 10), destacando comunidades como la **Comunidad Valenciana, Galicia, la Comunidad de Madrid y las Islas Baleares, donde se observan los mayores porcentajes de depresión**. En el extremo opuesto, las **menores proporciones se registran en Melilla, Extremadura y Ceuta**, lo que evidencia una variabilidad territorial relevante en la prevalencia de la depresión. Este patrón coincide con lo observado en otros estudios realizados en España, donde también se identifican diferencias territoriales en la salud mental (Henares, Ruiz-Pérez y Sordo, 2019). En particular, estos autores señalan a Galicia como una de las comunidades con mayor prevalencia, mientras que Ceuta y Melilla presentan los valores más bajos, en línea con los resultados obtenidos en este análisis. Estas variaciones pueden explicarse por factores como las condiciones económicas, las oportunidades laborales o el acceso a los servicios sanitarios, que no son homogéneos entre comunidades. Además, el carácter descentralizado del sistema sanitario español implica que cada comunidad autónoma gestiona y organiza sus propios recursos sanitarios (Álvarez, 2023, p. 118).

Gráfico 10: Porcentaje de depresión por Comunidad Autónoma



Fuente: Elaboración propia

4. IDENTIFICACIÓN Y TRATAMIENTO DE VALORES ATÍPICOS

Un valor atípico, también conocido como ***outlier***, hace referencia a una o más observaciones que se **desvían de manera significativa del resto de los datos** en la muestra (Barnett y Lewis, 1994). Es necesario reconocer y manejar estos valores o, en algunos casos, eliminarlos, porque tienen el potencial de impactar los resultados, obstaculizando la interpretación adecuada del análisis y sesgando las conclusiones posteriores.

No se trataron los valores atípicos en las variables con categorías predefinidas (por ejemplo, del 1 al 5 en la variable estado civil), ya que se considera que todas las respuestas se encuentran dentro del rango esperado y aportan información relevante para el análisis. En cambio, para las **variables numéricas**, edad, consumo medio diario de alcohol y peso, cuyas respuestas son de tipo "libre", sí se procedió a analizar la posible presencia de valores atípicos. Con el fin de detectar observaciones que se desvían notablemente del comportamiento general de la muestra, se utilizó el **método del rango intercuartílico (IQR).** Este procedimiento, que se apoya en la posición relativa de los datos dentro de la distribución, posibilita identificar valores extremos de manera objetiva y estadísticamente fundamentada.

El rango intercuartílico (IQR) se calcula como la **diferencia entre el tercer cuartil (Q3) y el primer cuartil (Q1)**, es decir, IQR = Q3 – Q1. Un dato se considera atípico si se **encuentra fuera del intervalo definido** por [Q1 – 1.5*IQR, Q3 + 1.5*IQR]. En otras palabras, cualquier valor que esté más de 1,5 veces el IQR por debajo de Q1 o por encima de Q3 se clasifica como un valor atípico y, por tanto, puede ser eliminado o tratado según el caso (Ministerio de Asuntos Económicos y Transformación Digital, 2021).

Se detectaron valores atípicos únicamente en dos de las variables numéricas estudiadas: **el peso, con 376 casos, y el consumo medio diario de alcohol, con 2.682 casos**. El impacto del tratamiento de *outliers* en la media, la dispersión y el rango de las variables analizadas se puede observar a partir de los estadísticos descriptivos que presenta la Tabla 7.

Tabla 7: Estadísticos del consumo medio de alcohol y del peso con y sin outliers

Variable	Dataset	Media	Desv. estándar	Mínimo	P25	P75	Máximo
Consumo medio alcohol (g)	Original	3,95	8,86	0	0	4,29	240
	Sin atípicos	1,36	2,87	0	0	0	10
Peso (kg)	Original	72,96	14,32	33	63	81	180
	Sin atípicos	71,36	12,82	35	62	80	107

Fuente: Elaboración propia

Antes de ser depuradas, las dos mostraban observaciones extremas. El consumo medio diario de alcohol alcanzaba las 240 unidades y el peso, por su parte, llegaba hasta los 180 kg. Después de eliminar los *outliers*, estos valores se disminuyeron a 10 y 107 respectivamente, quedando dentro de rangos más acordes con el comportamiento general observado en la muestra. Las medidas de dispersión y tendencia central también se vieron impactadas por la depuración. En el consumo medio diario de alcohol, la media y la desviación estándar disminuyeron de forma considerable, lo que refleja una reducción de la variabilidad y una distribución más homogénea. En cuanto al peso, las dos mediciones también disminuyeron, pero de forma más moderada, y no hubo alteraciones significativas en la tendencia central de los datos.

Eliminar los valores atípicos en el consumo medio diario de alcohol reduciría el valor máximo a 10 gramos, reflejando solo patrones de consumo moderado. Sin embargo, dado que el consumo mundial medio ronda los 33 gramos diarios (OMS, 2018) y a partir de 60 en hombres y 40 en mujeres se considera de riesgo (Ministerio de Sanidad y Consumo, 2008), excluir los 48 casos que superan este umbral implicaría perder información relevante sobre posibles comportamientos de abuso o dependencia, fundamentales para los objetivos del estudio.

De manera similar, en la variable peso, los valores muy altos podrían pertenecer a individuos con obesidad severa, un subgrupo de interés en el análisis de la depresión. Se decidió, por tanto, también conservar estas observaciones para analizarlas más adelante.

5. TRATAMIENTO DE VALORES PERDIDOS

La Tabla 8 presenta un resumen de los 35.925 valores NaN (perdidos) que se detectaron en las diferentes variables del conjunto de datos.

Tabla 8: Número y porcentaje de valores perdidos por variable

Variable	Valores perdidos	Valores perdidos (%)
SEVERIDAD_DEPRESIVA	699	3,32
consumo_medio_diario_alcohol	51	0,24
apoyo_social_percibido	287	1,36
Fumador	113	0,54
actividad_física	271	1,29
Peso	867	4,12
consumo_tranquilizantes	7701	36,62
nivel_estres_laboral	11298	53,72
jornada_laboral	11248	53,48
estado_civil	135	0,64
convivencia_en_pareja	395	1,88
Sexo	0	0
Edad	0	0
CCAA	1802	8,57
lesiones_permanentes_accidente	91	0,43
tumor_maligno	190	0,90

Fuente: Elaboración propia

En primer lugar, se **eliminaron** los registros que contenían va**lores perdidos en la variable objetivo** porque la falta de datos en la variable a predecir podría generar sesgos en las predicciones. Asimismo, se decidió **eliminar las variables con más del 50 % de valores perdidos,** en concreto la **jornada de trabajo y el nivel de estrés laboral**, ya que su alto porcentaje de datos ausentes impedía una imputación fiable.

5.1 Imputación simple

Para abordar los valores perdidos en las **variables categóricas nominales y ordinales** que tienen un **porcentaje inferior al 5% de valores perdidos**, se aplicó una estrategia de imputación basada en la **moda**, es decir, se sustituyó cada valor ausente por la categoría más frecuente dentro de la misma variable. Estas variables están codificadas numéricamente, aunque sean categóricas lo que justifica el uso de la moda y que el resultado salga en número.

Las imputaciones realizadas para las variables categóricas fueron las siguientes:

- **estado_civil** → imputado con 2 (categoría más frecuente: casado/a)
- **tumor_maligno**, **lesiones_permanentes_accidente** → imputadas con 0 (categoría: no)

- **convivencia_en_pareja** → imputado con 0 (categoría: no convive en pareja)
- **actividad_física** → imputado con 2 (categoría: actividad física ocasional)
- **fumador** → imputado con 3 (categoría: no, nunca ha fumado)
- **apoyo_social_percibido** → imputado con 4 (categoría: más de cinco personas)

En el caso de las **variables numéricas** con **menos del 5 % de datos faltantes**, se aplicó una estrategia de imputación basada en la **media aritmética**, una técnica común cuando se busca preservar la tendencia central de la variable sin introducir sesgos importantes.

Las imputaciones realizadas fueron las siguientes:

- **peso** → imputado con una media de 72,97 kg
- **consumo_medio_diario_alcohol** → imputado con una media de 3,98 unidades

5.2 Imputación múltiple

En el conjunto de datos analizado, únicamente **CCAA** (8,6%) y **consumo_tranquilizantes** (36,99%) cuentan con un porcentaje de valores perdidos **superior al 5%.** Para estas variables se aplicó un proceso de **imputación múltiple,** una metodología estadística avanzada que permite reconstruir la información faltante de manera más precisa y realista que las imputaciones simples. A diferencia de la imputación simple, que reemplaza cada dato perdido por un único valor estimado como la media o la moda, la imputación múltiple **genera varias versiones completas del conjunto de datos,** en las que los **valores ausentes se sustituyen por estimaciones diferentes obtenidas mediante procedimientos aleatorios** (Pineda, 2025). Una vez generadas todas las versiones del conjunto de datos, se combinan para formar un único conjunto final que integra la información obtenida en cada una de ellas, garantizando así resultados más consistentes y representativos (Scikit-learn Developers, 2024).

En este estudio se generaron **20 imputaciones independientes**, cada una con una **semilla aleatoria diferente** (semillas 0 a 19), para garantizar variabilidad entre versiones. Dado que CCAA y consumo_tranquilizantes presentan valores perdidos simultáneamente y pueden estar relacionadas entre sí, las variables se imputaron de forma alternada en **5 ciclos sucesivos.** En el primer ciclo, se imputa CCAA utilizando consumo_tranquilizantes como predictor; a continuación, se imputa consumo_tranquilizantes utilizando la CCAA

recién estimada. En el segundo ciclo se repite el mismo proceso, pero ahora con las estimaciones mejoradas del ciclo anterior, y así sucesivamente hasta completar los 5 ciclos. Este proceso iterativo permite que ambas variables se vayan refinando mutuamente hasta alcanzar estimaciones estables.

Al tratarse de variables **categóricas nominales**, se empleó un **modelo de Random Forest con muestreo probabilístico**, configurado con **300 árboles de decisión** (n_estimators = 300) y **profundidad máxima libre** (max_depth = None), lo que permite que cada árbol capture relaciones complejas entre las variables sin restricciones de tamaño. Por ejemplo, para imputar los valores perdidos de consumo_tranquilizantes, el modelo se entrena con los individuos que sí tienen ese dato, utilizando CCAA y el resto de variables como predictores, e identifica qué patrones se asocian con consumir o no tranquilizantes. Una vez entrenado, estima la probabilidad de que cada observación pertenezca a cada categoría posible, en este caso, consumidor o no consumidor, y asigna el valor ausente de forma aleatoria pero ponderada según dichas probabilidades, respetando la incertidumbre inherente a la imputación. El mismo proceso se aplica a CCAA, estimando la comunidad autónoma más probable para cada individuo con ese dato perdido. Para obtener un único conjunto de datos final, se aplicó la **moda por fila**: a cada observación se le asignó la categoría que apareció con mayor frecuencia entre las 20 imputaciones.

CAPÍTULO III: SELECCIÓN DE VARIABLES

1. EXPLICACIÓN Y JUSTIFICACIÓN DEL USO DE *MUTUAL INFORMATION*

La **información mutua (MI)** es una medida de la teoría de la información que permite **cuantificar la relación entre dos variables y conocer en qué medida el valor de una ayuda a predecir la otra** (Witten et al., 2016). Su utilidad en ML se basa en que permite identificar y seleccionar únicamente las **variables más relevantes e informativas** para el modelo, descartando aquellas que no contribuyen de forma significativa a la predicción (Suvendu, 2024). Esto es particularmente relevante cuando se manejan numerosas variables, pues el exceso de variables que no tienen relevancia puede dar lugar a un sobreajuste (*overfitting*): el modelo memoriza los datos de entrenamiento, incluido el ruido, y presenta un bajo desempeño al ser utilizado con nuevos casos (MathWorks, s.f.).

Desde un punto de vista matemático, la información mutua entre dos variables X e Y se expresa como (Walters-Williams y Li, 2009):

$$I(X; Y) = H(X) - H(X|Y)$$

En esta expresión, I (X; Y) simboliza la información mutua entre las dos variables, H(X) alude a la variabilidad o incertidumbre inicial de X en ausencia de otra información, y H(X|Y) representa la incertidumbre que persiste en X cuando ya se ha obtenido Y, lo que se denomina entropía condicional. En otras palabras, la información mutua **mide cuánta incertidumbre se reduce en una variable gracias a la información aportada por la otra** (Walters-Williams y Li, 2009).

En la práctica, el método implica calcular el valor de información mutua de cada variable predictora en relación con la variable objetivo. Los valores de MI, que se representan en **bits**, siempre son mayores o iguales a cero. Si dos variables son **independientes**, saber una no proporciona ninguna pista acerca de la otra, en consecuencia, la incertidumbre permanece igual y el valor de la información mutua es **0.** A medida que existe mayor dependencia entre ellas, la incertidumbre se reduce más y la información mutua es más elevada, reflejando una relación relevante (Data Science Python Blog, 2023; Zhu, 2021).

La principal ventaja de MI en comparación con otros métodos más simples para seleccionar variables, como PCA o Chi-cuadrado, es que **puede detectar cualquier forma de dependencia entre las variables, no solo las lineales**. Esto la hace una técnica

más sólida y capaz de percibir patrones reales que están presentes en los datos (Barañao, s.f.).

2. RESULTADOS Y SELECCIÓN DE VARIABLES

Para la selección de variables se empleó el estimador **mutual_info_classif** de la librería scikit-learn, que permite calcular la información mutua entre cada predictor y una **variable objetivo discreta** (Nair, 2023). Dado que esta función emplea procedimientos diferentes para las variables discretas y continuas, fue preciso establecer de antemano una **máscara** que determinara cuáles predictores debían ser considerados como continuos y cuáles como categóricos o discretos.

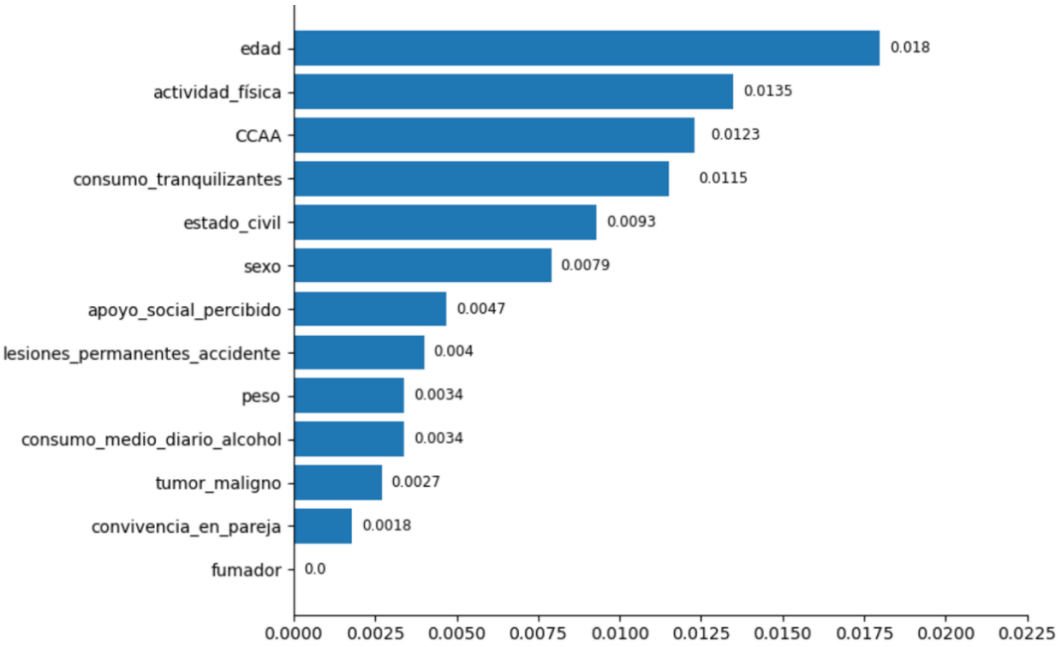
El cálculo de la información mutua presenta cierta variabilidad debido a componentes aleatorios del estimador. Con el fin de conseguir resultados consistentes y prevenir que dependieran de una sola ejecución, se realizó el **procedimiento 1000 veces** y se calculó la **media** para determinar el valor final. Esta perspectiva ofrece una estimación más sólida de la aportación informativa que cada variable tiene.

Un valor de **información mutua mayor que 0,10** se considera en la literatura como un indicador de una contribución informativa moderada (Oviedo, 2025). Ninguna de las variables analizadas supera dicho umbral, lo cual es coherente con la naturaleza multifactorial de la depresión (Ver Gráfico 11 y Tabla 9). La aparición de un cuadro depresivo no puede ser explicada por un solo predictor, sino que es el resultado de la interacción de diferentes dimensiones sociales, psicológicas y biológicas (Remes, Mendes y Templeton, 2021). Aun así, algunas variables destacan por tener una contribución relativa más alta, las que más información aportan son **edad** (MI = 0,0180), **actividad_física** (MI = 0,0135), **CCAA** (MI = 0,01123) y **consumo_tranquilizantes** (MI = 0,0115).

La relevancia de la actividad física es consistente con la evidencia empírica previa, que ha demostrado de forma reiterada su relación con la salud mental. En particular, estudios como los de Matute-Reyes y Torres-Palchisaca (2025) y Agudelo, Ante y Torres (2017) concluyen que niveles elevados de actividad física se asocian con una menor probabilidad de desarrollar trastornos depresivos, actuando como un factor protector. Asimismo, la

edad y el lugar de residencia (representado en este estudio a través de la comunidad autónoma) también han sido identificados como variables relevantes en el análisis de la salud mental. Moreno, Lostao y Regidor (2023) evidencian que los problemas de salud mental presentan variaciones significativas tanto entre distintos grupos de edad como entre territorios, lo que refleja la influencia de factores demográficos y contextuales en la prevalencia de la depresión.

Gráfico 11: Importancia de las variables explicativas de la severidad depresiva según información mutua



Fuente: Elaboración propia

Tabla 9: Importancia de las variables explicativas de la severidad depresiva según información mutua

Variable	Información mutua media
Edad	0,0180
Actividad_fisica	0,0135
CCAA	0,0123
consumo_tranquilizantes	0,0115
Estado_civil	0,0093
Sexo	0,0079
Apoyo_social_percibido	0,0047
Lesiones_permanentes_accidente	0,0040
Consumo_medio_alcohol	0,0034
Peso	0,0034
Tumor_maligno	0,0027
Convivencia en pareja	0,0018
Fumador	0,0000

Fuente: Elaboración propia

CAPÍTULO IV: MODELOS DE MACHINE LEARNING

El propósito de este capítulo es crear y evaluar diferentes modelos de *machine learning* para determinar cuál tiene un rendimiento superior en la predicción de la depresión en los adultos españoles. El **aprendizaje automático** (*machine learning*) consiste en desarrollar **algoritmos** que tienen la **capacidad de aprender patrones** con los datos existentes y **optimizar su desempeño** a medida que integran información nueva, creando modelos capaces de hacer **predicciones sobre observaciones previamente no observadas** (Zhou, 2021, p. 2). Los modelos se entrenan utilizando como predictores aquellas variables que presentaron los valores más elevados de información mutua con la depresión (edad, actividad_física, CCAA y consumo_tranquilizantes).

La variable objetivo de este estudio muestra un desbalance entre las clases: cerca del 32 % de los participantes de la muestra tiene depresión, mientras que el 69% no la tiene (Ver Gráfico 4). Para una **proporción de 70/30, no se considera necesario el empleo de técnicas particulares para datos desbalanceados.**

Para evaluar si los modelos aportan valor real, se incluye un **modelo trivial (*Dummy Classifier*)** como referencia (Scikit-learn developers, s.f.; Sahel, s.f.). Este clasificador **predice siempre la clase más frecuente** en los datos de entrenamiento, en este caso, ausencia de depresión, sin aprender ningún patrón. Al hacerlo, acierta automáticamente en todos los casos de la clase mayoritaria y falla en todos los de la clase minoritaria, **obteniendo una exactitud del 69%,** que coincide exactamente con la proporción de personas sin depresión en la muestra. Todo modelo de *machine learning* desarrollado en este estudio debe obtener resultados notablemente mejores que el *Dummy Classifier*.

Las métricas empleadas para evaluar los diferentes modelos se resumen en la siguiente tabla:

Tabla 10: Métricas de evaluación de los modelos

Métrica	Definición	Fórmula
Exactitud (Accuracy)	Proporción de todas las clasificaciones que fueron correctas, ya sean positivas o negativas	$\frac{TP+TN}{TP+TN+FP+FN}$
Recall (Sensibilidad o Tasa de verdaderos positivos)	Proporción de todos los positivos reales que se clasificaron correctamente como positivos	$\frac{TP}{TP+FN}$
Precisión	Proporción de todas las clasificaciones positivas del modelo que son realmente positivas.	$\frac{TP}{TP+FP}$

F1-score	Es la media de la precisión y sensibilidad. El F1-score solo logra altos valores cuando ambos elementos tienen un buen rendimiento. Por lo tanto, si el modelo reduce su capacidad para detectar correctamente los casos positivos (*recall*) o disminuye la precisión de tales predicciones, esta limitación se traduce inmediatamente en una caída del valor del indicador	$2 * \frac{Precisión * Recall}{Precisión + Recall}$
AUC-ROC (Área Bajo la Curva)	Mide la capacidad del modelo para distinguir correctamente entre casos con y sin depresión, con independencia del umbral de clasificación. Toma valores entre 0 y 1. Un valor de 1 indica que el modelo distingue todas las clases sin error.	

Fuente: Google Developers (s.f.)

1. MODELOS APLICADOS

1.1 Regresión Logística

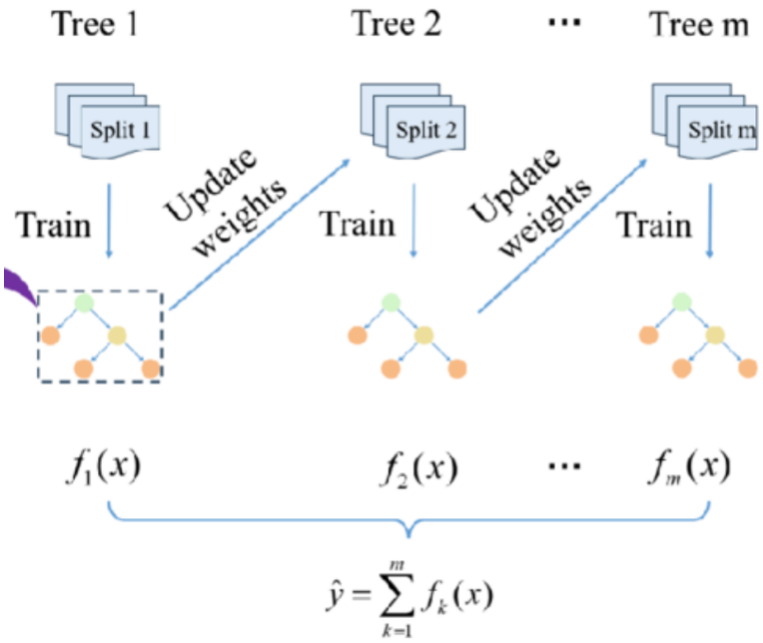
La regresión logística es un algoritmo de aprendizaje automático supervisado utilizado para problemas de clasificación. Este modelo analiza la relación entre las variables predictoras (variables independientes) y una variable de salida u objetivo (variable dependiente). Se utiliza cuando la variable objetivo es categórica, como ocurre en este estudio, donde la variable objetivo es la presencia de depresión. El modelo **estima la probabilidad de que ocurra este resultado en función de un conjunto de variables consideradas relevantes o influyentes**. A partir de dicha probabilidad, el modelo clasifica cada caso en la categoría correspondiente (Guillén y Alonso, 2020).

1.2 *XGBoost*

El algoritmo de aprendizaje automático supervisado conocido como *XGBoost* (*Extreme Gradient Boosting*) se basa en **árboles de decisión**. Su funcionamiento se basa en la idea de combinar múltiples árboles de forma secuencial a fin de incrementar gradualmente su capacidad predictiva. **Cada árbol nuevo se entrena con los resultados obtenidos por los árboles anteriores para corregir los errores**. Este procedimiento sigue ocurriendo **hasta que el modelo no puede disminuir más el error** (gradiente descendiente) (Espinosa-Zúñiga, 2020). Además, cada árbol contribuye de manera distinta al resultado final porque sus predicciones se ponderan a través de una tasa de aprendizaje. La predicción final se logra al agregar las salidas de todos los árboles, ponderadas por la tasa (Espinosa-Zúñiga, 2020). Las ventajas más significativas de *XGBoost* son su rapidez de ejecución y su capacidad para predecir (Mitchell y Frank, 2017). Además, este algoritmo es eficaz con conjuntos de datos grandes y complejos, como el que se presenta en esta

investigación (Esri, s.f.).

Ilustración 1: Representación del modelo XGBoost



Fuente: (Xiong, Chen., Zhang, et al.) 2024

2. COMPARACIÓN DE RESULTADOS

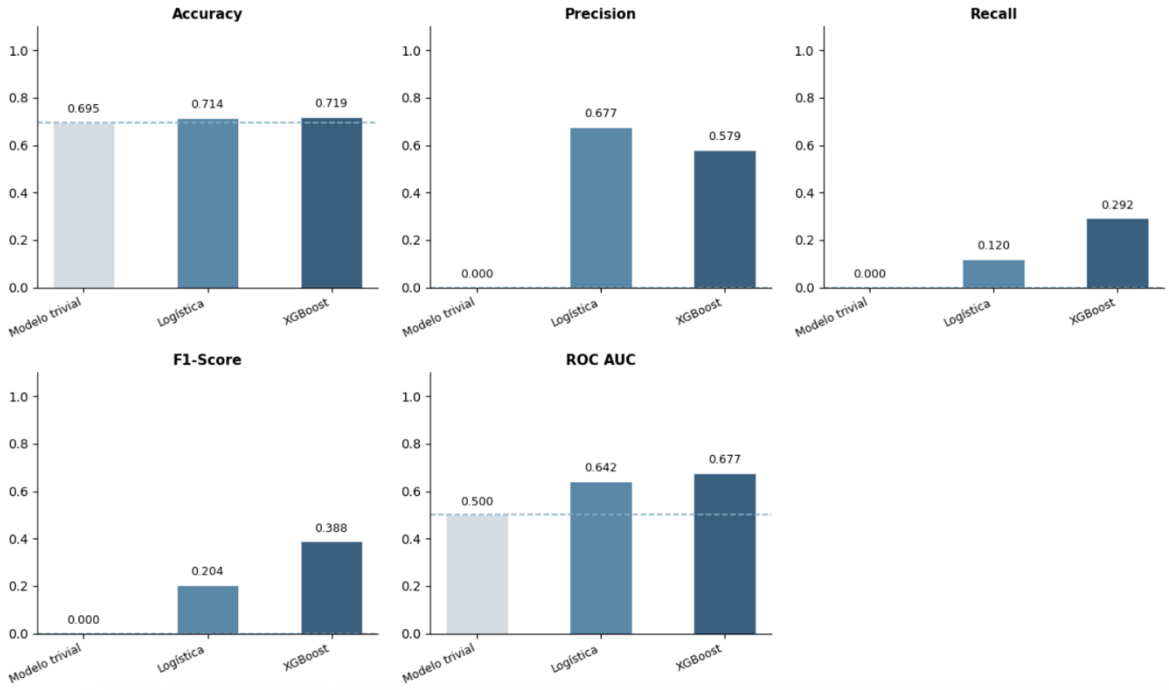
Los resultados obtenidos por los diferentes modelos se presentan en la Tabla 11, Gráfico 12 y 13. Cabe señalar que las métricas y la matriz de confusión se calcularon de forma acumulada sobre las 20 imputaciones, por lo que el número total de casos reflejado en la matriz es superior al tamaño de la muestra original.

Tabla 11: Desempeño comparativo de los modelos de clasificación según las métricas de Exactitud, Precisión de la clase positiva, Sensibilidad, F1-score, AUC y la matriz de confusión expresada como [TN, FP; FN, TP]

Modelo	Exactitud (accuracy)	Precisión (1)	Sensibilidad (Recall) (1)	F1-Score	AUC	Matriz de Confusión
Modelo trivial	0,6948	0	0	0	0,5	
Logística	0,7139	0,6772	0,1201	0,2040	0,6420	[56983; 1477; 22595; 3085]
XGBoost	**0,7190**	**0,5789**	**0,2915**	**0,3877**	**0,6766**	**[53009; 5451; 18193; 7487]**

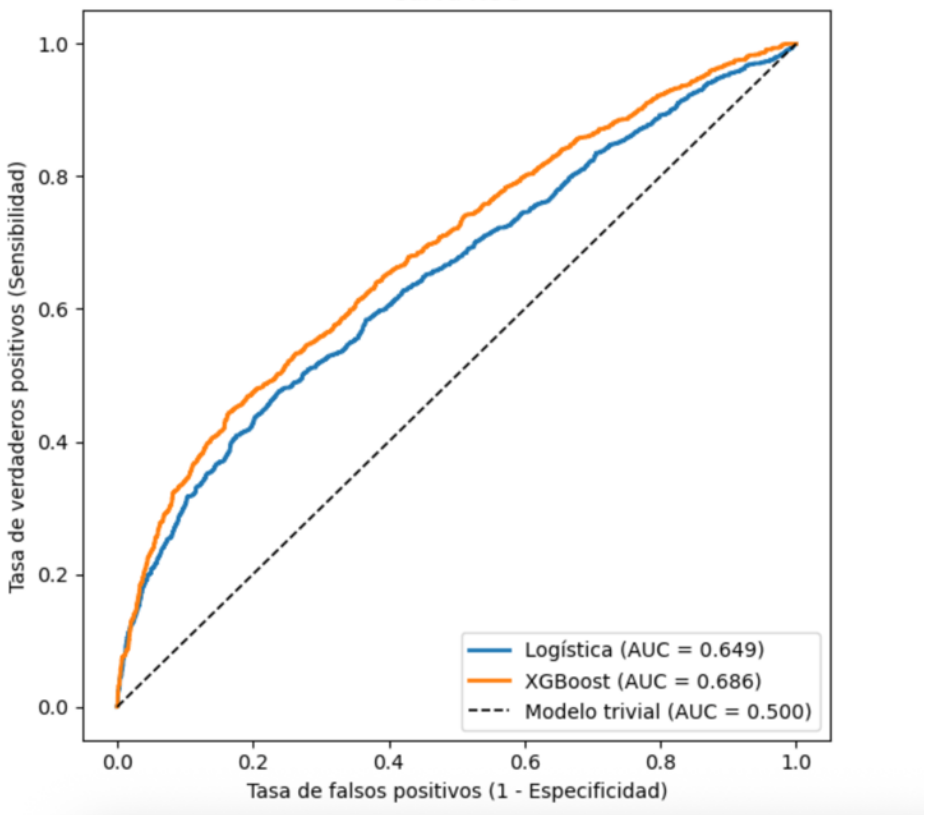
Fuente: Elaboración propia

Gráfico 12: Comparativa de métricas de evaluación por modelo



Fuente: Elaboración propia

Gráfico 13: Curva ROC comparativa entre la regresión logística y XGBoost



Fuente: Elaboración propia

El modelo trivial obtiene una exactitud del 69,5 % prediciendo siempre la ausencia de depresión, con precisión, sensibilidad y F1-Score iguales a 0 y un AUC de 0,500,

confirmando que carece de capacidad discriminativa y actúa únicamente como umbral mínimo de comparación.

La **regresión logística** alcanza una exactitud del 71,4 % y una precisión del 67,7 %, lo que indica que la mayoría de las personas clasificadas como deprimidas realmente lo están. Sin embargo, su sensibilidad es muy reducida (12,0 %), ya que solo identifica correctamente 12 de cada 100 casos reales de depresión, dejando un elevado número de falsos negativos (22.595). Este desequilibrio entre precisión y sensibilidad se refleja en un F1-Score bajo (0,204). Por su parte, el AUC de 0,642 muestra una capacidad discriminativa moderada, superior al clasificador aleatorio.

XGBoost presenta el mejor desempeño global. Aunque su precisión es algo inferior (57,9 %), mejora notablemente la sensibilidad hasta el 29,2 %, detectando correctamente un número mucho mayor de personas con depresión (7.487 verdaderos positivos frente a 3.085 en la regresión logística) y reduciendo los falsos negativos. Este mejor equilibrio entre precisión y capacidad de detección se traduce en un F1-Score superior (0,388) y en el mayor AUC (0,677), indicando una mejor capacidad para distinguir entre personas con y sin depresión. Por ello, XGBoost se selecciona como modelo de referencia para el análisis de interpretabilidad del capítulo siguiente. Otros estudios centrados en la predicción de síntomas depresivos, como el de Lu, Wan y Liu (2025), muestran que *XGBoost* fue el modelo con mejor desempeño en una base de datos también desbalanceada, en la que la presencia de depresión constituía la clase minoritaria, al igual que en este estudio. Asimismo, el mismo estudio constató que el algoritmo *LightGBM* también presentó un rendimiento elevado.

CAPÍTULO V: MODELO INTERPRETABLE

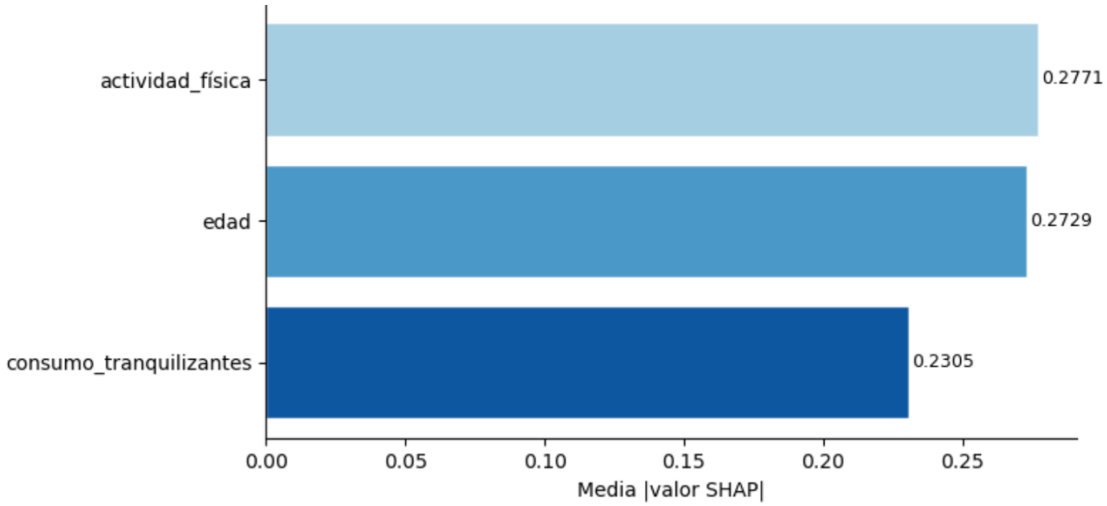
En los problemas de clasificación, es común que los modelos más efectivos en términos de predicción sean de **tipo caja negra**, como sucede con algoritmos basados en árboles de decisión, siendo *XGBoost* un ejemplo significativo. A pesar de que estos algoritmos generalmente logran grados altos de precisión, su estructura interna es compleja y poco transparente, lo cual complica comprender cómo las diferentes variables afectan las predicciones generadas por el modelo (Lundberg y Lee, 2017).

Por esta escasa capacidad de interpretación, se **empleó SHAP (*SHapley Additive exPlanations*),** un método que permite **comprender las predicciones realizadas por modelos complejos.** En este escenario, el análisis se realizó sobre el modelo XGBoost, ya que mostró la mayor capacidad predictiva en el capítulo anterior. Esta técnica permite descomponer cada predicción del modelo en contribuciones individuales de las variables explicativas. El uso de SHAP resulta especialmente relevante cuando el objetivo no es únicamente realizar predicciones, sino también comprender qué factores influyen en la variable objetivo. En el contexto de este estudio, esta metodología permite analizar qué variables están más asociadas con un mayor riesgo de padecer depresión, facilitando así una interpretación más clara del comportamiento del modelo. A través de esta técnica fue posible obtener interpretaciones tanto globales como individuales de los resultados del modelo XGBoost.

Aunque el modelo XGBoost fue entrenado con cuatro variables, **la variable CCAA fue excluida** del análisis SHAP al tratarse de una variable nominal sin orden intrínseco. A diferencia de variables como la edad o la actividad física, donde un valor más alto implica sistemáticamente un mayor o menor riesgo, las categorías de CCAA no siguen una escala direccional, lo que impide una interpretación coherente de sus valores SHAP.

El Gráfico 14 recoge la importancia media de cada variable medida como el valor absoluto medio de sus contribuciones SHAP. Las tres variables presentan una importancia similar: **actividad_física (0,277), edad (0,273) y consumo de tranquilizantes (0,231),** lo que señala que las tres contribuyen de manera equilibrada a explicar la probabilidad de depresión predicha por el modelo.

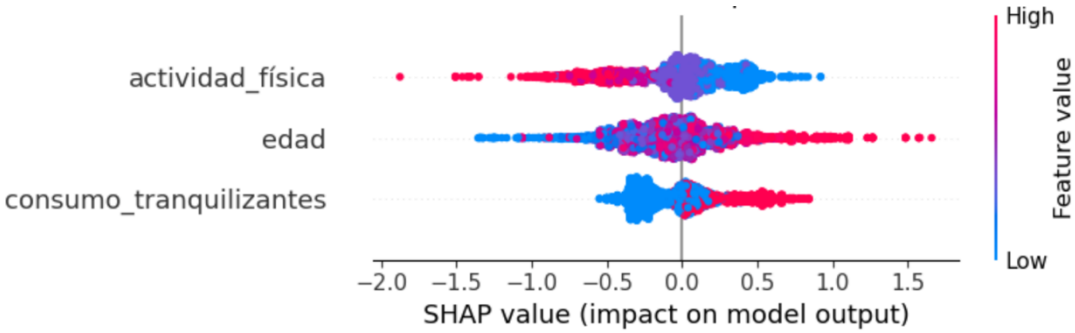
Gráfico 14: Importancia global de las variables según SHAP — XGBoost



Fuente: Elaboración propia

El Gráfico 15 permite interpretar no solo la **importancia de cada variable**, sino también la **dirección de su efecto**. Cada punto corresponde a un individuo de la muestra, y su ubicación en el eje horizontal muestra cómo influye esa variable en el resultado del modelo. Los valores SHAP **positivos** indican que la variable **contribuye a aumentar** la probabilidad estimada de depresión, mientras que los **valores negativos** indican que **reduce dicha probabilidad** o la orienta hacia la ausencia de depresión. Por su parte, el color de los puntos permite interpretar la intensidad o nivel de la variable: los tonos rosas representan valores elevados y los tonos azules valores bajos.

Gráfico 15: Diagrama de dispersión SHAP (beeswarm)



Fuente: Elaboración propia

A partir del gráfico se observa que **actividad_física es la variable con mayor peso** en el modelo, ya que aparece en la primera posición y presenta la mayor dispersión de valores SHAP. Se observa que los valores altos (rosa) tienden a situarse en la zona negativa del eje, lo que implica que **niveles elevados de actividad física reducen la probabilidad** de depresión. Por el contrario, los valores bajos (azul) se concentran en la zona positiva,

aumentando dicha probabilidad. Este resultado sugiere un claro **efecto protector de la actividad física.**

En segundo lugar, en cuanto a la **edad**, los valores altos (rosa) se sitúan principalmente en la zona positiva del eje, lo que indica que una **mayor edad incrementa la probabilidad** de depresión según el modelo. Por el contrario, los valores bajos (azul) tienden a reducir dicha probabilidad. Esto sugiere que el **riesgo de depresión aumenta con la edad** en el conjunto de datos analizado.

Por último, el **consumo de tranquilizantes** también tiene un impacto significativo. Los valores altos se asocian con contribuciones positivas al modelo, lo que indica que el **consumo de este tipo de medicamentos aumenta la probabilidad** predicha de depresión. Esto puede interpretarse como un indicador indirecto de problemas de salud mental previos o coexistentes.

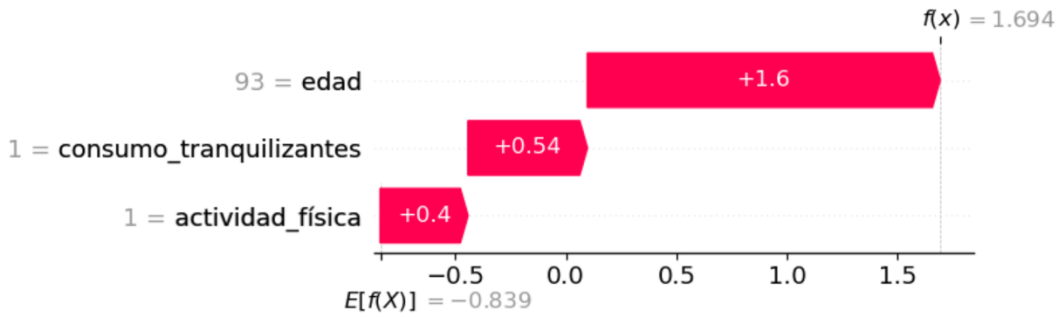
Por tanto, los resultados sugieren que niveles bajos de actividad física, una mayor edad y el consumo de tranquilizantes son factores que incrementan la probabilidad predicha de depresión. En particular, la actividad física actúa como un factor protector, mientras que la edad y el consumo de tranquilizantes se asocian con un mayor riesgo. Estos resultados son consistentes con la literatura previa, que ha evidenciado una relación inversa entre la actividad física y la depresión (Matute-Reyes y Torres-Palchisaca, 2025). Asimismo, diversos estudios han señalado que la probabilidad de presentar problemas de salud mental aumenta con la edad (Moreno, Lostao y Regidor, 2023).

Posteriormente, con el objetivo de ilustrar cómo el modelo genera predicciones individuales, se emplearon gráficos *waterfall.* El punto de partida es el **valor base E[f(X)],** que representa la **predicción media** que haría el modelo si no se dispusiera de información sobre el individuo, es decir, la probabilidad media de depresión en la muestra. A partir de ahí, **cada variable suma o resta su contribución** de forma acumulativa (representada mediante barras rojas cuando aumentan la probabilidad predicha y azules cuando la reducen) hasta **alcanzar la predicción final f(x),** que es el **valor** concreto asignado a ese **individuo** por el modelo. Para ilustrar casos extremos, se seleccionaron dos individuos: uno procedente de la comunidad autónoma con mayor prevalencia de depresión en la muestra y otro de la comunidad con menor prevalencia,

con el fin de contrastar perfiles de riesgo opuestos y mostrar cómo las mismas variables operan en sentidos contrarios según las características del individuo.

El Gráfico 16 muestra **el perfil de alto riesgo**, correspondiente a un individuo de 93 años residente en la **Comunidad Valenciana**, sedentario y consumidor de tranquilizantes. El modelo parte del valor base **E[f(X)] = −0,839**. A partir de este punto, cada variable suma su contribución de forma acumulativa hasta alcanzar la predicción final del individuo analizado f**(x)=1,694** equivalente a una **probabilidad predicha de depresión del 84,5 %.** En este caso, la **edad** de 93 años constituye el **principal factor explicativo**, con una contribución de **+1,6**, incrementando de forma muy notable la probabilidad predicha. En segundo lugar, el **consumo de tranquilizantes** añade una contribución de **+0,54,** reforzando la asociación entre el uso de este tipo de medicamentos y una mayor probabilidad de depresión. Por último, el hecho de ser **sedentario** contribuye con **+0,4**, indicando que la falta de actividad física aumenta adicionalmente el riesgo.

Gráfico 16: SHAP waterfall — perfil de alto riesgo (Comunidad Valenciana)

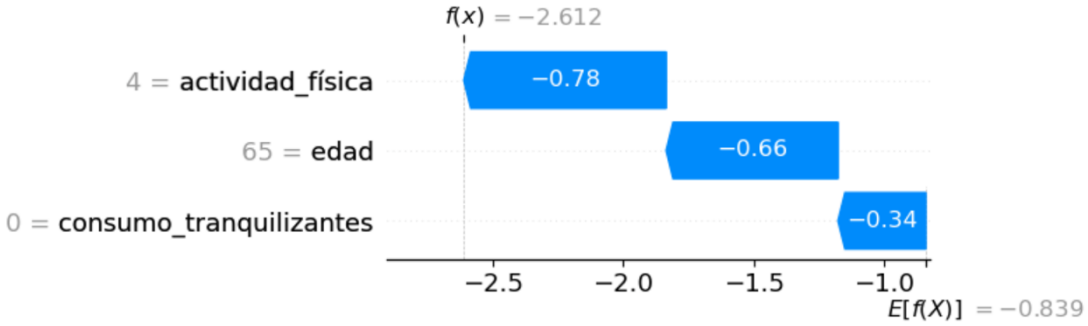


Fuente: Elaboración propia

El Gráfico 17 muestra el **perfil de bajo riesgo**, correspondiente a un individuo de 65 años residente en **Ceuta**, con actividad física intensa y sin consumo de tranquilizantes. Partiendo del mismo valor base que en el caso anterior (E[f(X)] = −0,839), todas las variables actúan aquí en sentido contrario, reduciendo la probabilidad predicha de forma acumulativa. El principal **factor protector** es el nivel muy elevado de **actividad física** (actividad_física = 4), que contribuye con **−0,78,** lo que indica que practicar actividad física de forma intensa reduce considerablemente la probabilidad de depresión predicha por el modelo. En segundo lugar, la **edad** de 65 años añade una contribución **de −0,66**: a diferencia del perfil de alto riesgo, una edad más moderada actúa en este caso como factor de protección. Por último, la **ausencia de consumo de tranquilizantes suma −0,34**,

reflejando que no tomar este tipo de medicamentos se asocia con una menor probabilidad de depresión. La suma acumulada de estas tres contribuciones negativas lleva la predicción final a **f(x) = −2,612**, equivalente a una **probabilidad predicha del 6,8 %.**

Gráfico 17: SHAP waterfall — perfil de bajo riesgo (Ceuta)



Fuente: Elaboración propia

La comparación entre ambos perfiles ilustra con claridad cómo las mismas tres variables operan en sentidos opuestos: la edad avanzada, el sedentarismo y el consumo de tranquilizantes configuran el perfil de mayor riesgo, mientras que la actividad física intensa, una edad media y la ausencia de consumo de tranquilizantes caracterizan el perfil de menor riesgo.

CAPÍTULO VI: REGLAS DE ASOCIACIÓN

1. METODOLOGÍA Y MÉTRICAS

Con el objetivo de profundizar en la interpretabilidad de los resultados y complementar los modelos de *machine learning* desarrollados en el capítulo anterior, se ha aplicado una **técnica de minería de datos basada en reglas de asociación**. Según Sharma y Verma (2014), la minería de datos consiste en **extraer patrones ocultos** y conocimiento relevante a partir del **análisis de grandes volúmenes** de información. Dentro de este ámbito, los algoritmos de reglas de asociación permiten **identificar relaciones frecuentes** entre variables, **generando reglas** que vinculan un **conjunto de condiciones** (**antecedente**) con un **determinado resultado** (**consecuente**), de modo que cuando el antecedente aparece, el consecuente tiende a producirse con elevada frecuencia (Sharma y Verma, 2014). Aplicado al presente estudio, este enfoque permite detectar qué **combinaciones de factores de riesgo** se asocian con **mayor frecuencia con la depresión severa**, facilitando la identificación de **perfiles de mayor vulnerabilidad.**

Para la generación de reglas se ha utilizado el **algoritmo Apriori**, implementado mediante la librería **mlxtend** de Python. Este algoritmo requiere que todas las variables sean **binarias** (Amat, 2018), por lo que las variables originales de la encuesta fueron recodificadas del siguiente modo, seleccionadas en coherencia con los resultados obtenidos en los capítulos anteriores y la literatura previa: mujer (sexo = mujer), edad_mayor55 (edad ≥ 55 años), sedentario (actividad física baja), toma_tranquilizantes (consumo de tranquilizantes = sí), sin_pareja (convivencia en pareja = no).

Las reglas obtenidas se analizan a partir de tres métricas fundamentales que permiten evaluar su relevancia y calidad:

Tabla 12: Métricas de evaluación de reglas de asociación

Métrica	Definición
Soporte	Representa el **porcentaje de registros** del conjunto de datos en los que **aparecen conjuntamente las condiciones del antecedente** y el **resultado del consecuente**. Esta métrica refleja la **frecuencia** con la que dicha asociación está presente en la muestra.
Confianza	Representa la **probabilidad** de que **ocurra el consecuente** cuando el **antecedente ya está presente**.
Lift	Representa la **relación entre la confianza de la regla y la frecuencia** con la que el consecuente aparece de manera general en la muestra. Un valor de ***lift* superior a 1** indica que la asociación entre las variables del antecedente y el consecuente es **mayor** de la que cabría esperar de **forma aleatoria**. Cuanto más elevado sea este valor, mayor será la fuerza y relevancia de la asociación detectada.

Fuente: Mwiti (2026) y Raschka (2023)

El análisis se centró en la **depresión severa** como variable objetivo, definida como SEVERIDAD_DEPRESIVA = 2, correspondiente a los casos de depresión moderada-grave o grave. Dado que esta variable presenta una prevalencia reducida en la muestra (6 %), se estableció un **soporte mínimo del 1 %** y una **confianza mínima del 5 %,** umbrales más bajos de lo habitual para no descartar **patrones relevantes**, pero **poco frecuentes.**

Una vez preparados los datos, el algoritmo se ejecuta en **cuatro fases** (Mwiti, 2026):

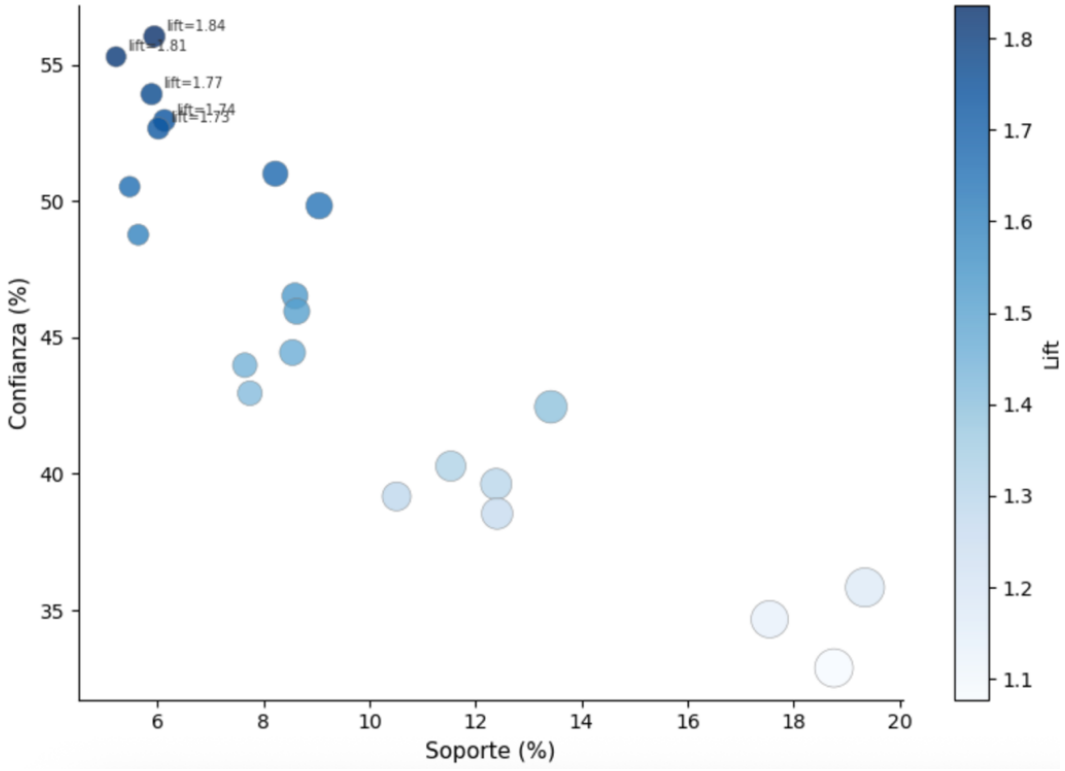
1. El algoritmo identifica las variables de forma **individual** y cuenta sus apariciones en la muestra, examinando con qué **frecuencia** aparecen características como ***sedentario***, *toma_**tranquilizantes*** o ***mujer*** entre los individuos analizados.
2. **Poda** según el **soporte mínimo**: se **eliminan** las combinaciones de variables cuya frecuencia de aparición **no supera el umbral del 1%** fijado, descartando así patrones demasiado **infrecuentes** para ser representativos.
3. **Generación de conjuntos de ítems frecuentes**: a partir de las variables frecuentes individuales, el algoritmo genera combinaciones de mayor tamaño de forma iterativa, por ejemplo, sedentario + toma_tranquilizantes hasta que no es posible formar nuevas combinaciones que superen el umbral de soporte.
4. A partir de los **conjuntos frecuentes** se extraen las **reglas cuyo consecuente es *dep_severa***, filtrando aquellas que **superan los umbrales de confianza y lift** establecidos, con el fin de identificar qué combinaciones de factores se asocian con mayor probabilidad a un cuadro depresivo grave. Con esta configuración, el algoritmo generó un total de 29 reglas.

2. RESULTADOS: REGLAS DE ASOCIACIÓN PARA DEPRESIÓN SEVERA

Como se indicó anteriormente, **se obtuvieron 29 reglas** con depresión severa como consecuente. El Gráfico 18 muestra un diagrama de dispersión en el que cada punto corresponde a una de ellas: el eje X recoge el soporte, el eje Y la confianza, el color el lift y el tamaño la frecuencia en la muestra. Se observa que las reglas con mayor lift y confianza se concentran en la zona superior izquierda, con soportes del 1,0–1,5 %, mientras que las más frecuentes (soporte 3,5–4,1 %) presentan confianzas más bajas (7–11 %) y lifts próximos a 1,5. Esto indica que los **perfiles de mayor riesgo son poco frecuentes**, pero **altamente discriminantes** cuando se presentan, lo que es consistente

con los modelos multifactoriales de la depresión, según los cuales los **cuadros más graves responden a la acumulación simultánea de múltiples factores de riesgo** más que a un único predictor dominante.

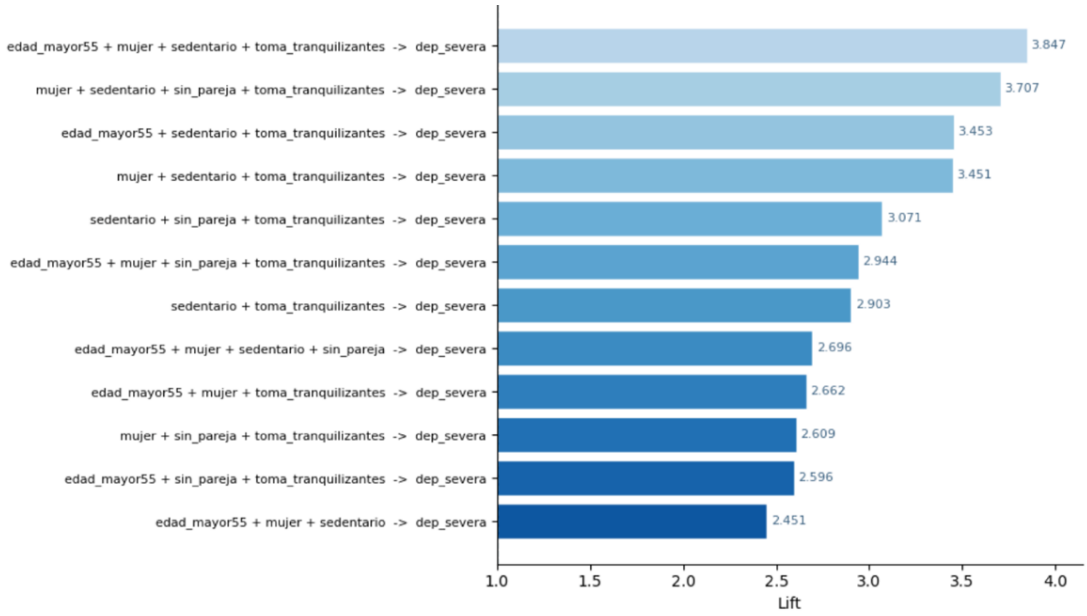
Gráfico 18: Diagrama de dispersión de las reglas de asociación para depresión severa según soporte, confianza y lift



Fuente: Elaboración propia

El Gráfico 19 presenta las 12 **reglas con mayor valor de lift**, ordenadas de forma descendente. Todas ellas muestran valores claramente **superiores a 1**, lo que indica que **no responden al azar**. Las cinco reglas con mayor lift superan el valor de 3, sugiriendo que los individuos con esos perfiles **presentan una probabilidad más de tres veces superior de padecer depresión severa** respecto al promedio de la muestra.

Gráfico 19: Reglas de asociación con mayor lift para depresión severa



Fuente: Elaboración propia

A continuación, se interpretan las reglas de asociación más relevantes:

Tabla 13: Reglas de asociación más relevantes para la identificación de depresión severa

Antecedente	Soporte	Confianza	Lift	Interpretación
edad_mayor55 + mujer + sedentario + toma_tranquilizantes	1,1%	23,1%	3,847	Se trata del **perfil de máximo riesgo**. Una mujer mayor de 55 años, sedentaria y consumidora de tranquilizantes tiene una **probabilidad estimada de depresión severa del 23,1 %, casi cuatro veces superior al 6 %** que presenta la población general de la muestra. Este perfil afecta al 1,1 % de la muestra, pero cuando se presenta el riesgo de depresión severa es muy elevado.
mujer + sedentario + sin_pareja + toma_tranquilizantes	1%	22,2%	3,707	Esta regla es prácticamente equivalente a la anterior en confianza y lift, con la diferencia de que sustituye edad_mayor55 por **sin_pareja**. Que ambas reglas alcancen valores tan similares sugiere que el **aislamiento social** tiene un peso comparable al del envejecimiento en la probabilidad de depresión severa, cuando se combina con sedentarismo y consumo de tranquilizantes.
edad_mayor55 + sedentario + toma_tranquilizantes	1,4%	20,7%	3,453	Esta regla, que **no incluye el sexo**, triplica igualmente el riesgo de depresión severa respecto al 6 % de la población general de la muestra (confianza = 20,7 %; lift = 3,453), lo que indica que **el perfil de riesgo no es exclusivo de las mujeres.** Su soporte del 1,4 % es ligeramente superior al de las dos reglas anteriores (1,1 % y 1,0 %), lo que significa que este patrón es algo **más frecuente** en la muestra. La comparación con la Regla 1, que sí incluye mujer y obtiene un lift de 3,847 frente al 3,453 de esta regla, sugiere que ser **mujer añade capacidad predictiva adicional,** pero **no es imprescindible** para configurar un perfil de alto riesgo.
mujer + sedentario + toma_tranquilizantes	1,5%	20,7%	3,451	Con una confianza del 20,7 % y un lift de 3,451, es prácticamente idéntica a la Regla 3 pero **sin restricción de edad**, lo que confirma que el **sedentarismo y el consumo de tranquilizantes configuran por sí solos un perfil de riesgo elevado en mujeres.** Su soporte del 1,5 %, el más alto de las cuatro, la convierte en el patrón más frecuente.

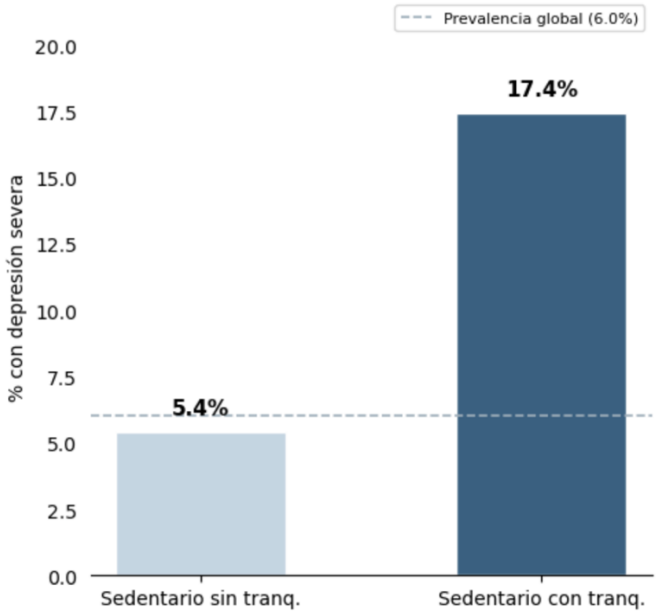
Fuente: Elaboración propia

Del conjunto de reglas obtenidas se extraen varias conclusiones relevantes. El **consumo de tranquilizantes** es el factor que **aparece en la mayoría** de las reglas con mayor lift, consolidándose como el **predictor individual más relevante**, resultado que coincide con el análisis SHAP. El **sedentarismo** está presente **en todas** las reglas de lift superior a 3, lo que evidencia su papel como **amplificador del riesgo** cuando se combina con otros factores. La **edad avanzada y el sexo femenino** aparecen de forma recurrente en los antecedentes, reflejando una **mayor vulnerabilidad** en mujeres a partir de los 55 años. Por último, **vivir sin pareja** destaca como **factor de riesgo** especialmente relevante en los cuadros severos, lo que sugiere que el **aislamiento social** adquiere mayor peso en los cuadros más graves de depresión.

3. CONTRASTE VISUAL DE LOS FACTORES DE RIESGO

Con el fin de contrastar visualmente los factores identificados en las reglas de asociación, se calculó la prevalencia de depresión severa comparando pares de grupos que difieren en un único factor de riesgo, manteniendo constante el resto.

Gráfico 20: Sedentario sin tranquilizantes vs Sedentario con tranquilizantes

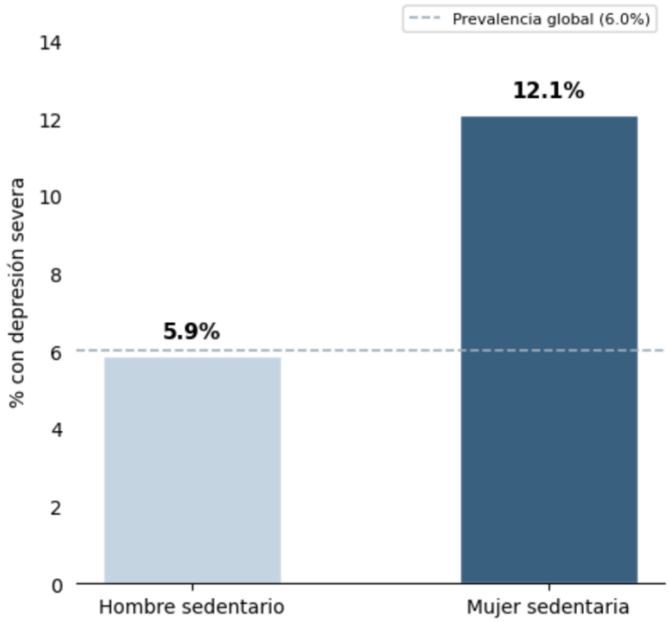


Fuente: Elaboración propia

El resultado más llamativo lo ofrece el Gráfico 20: entre los individuos sedentarios, quienes **consumen tranquilizantes** presentan una prevalencia del 17,4 % frente al 5,4 % de los que no los consumen. El consumo de tranquilizantes multiplica por más de tres el riesgo dentro del grupo sedentario, lo que explica su presencia en la totalidad de las reglas

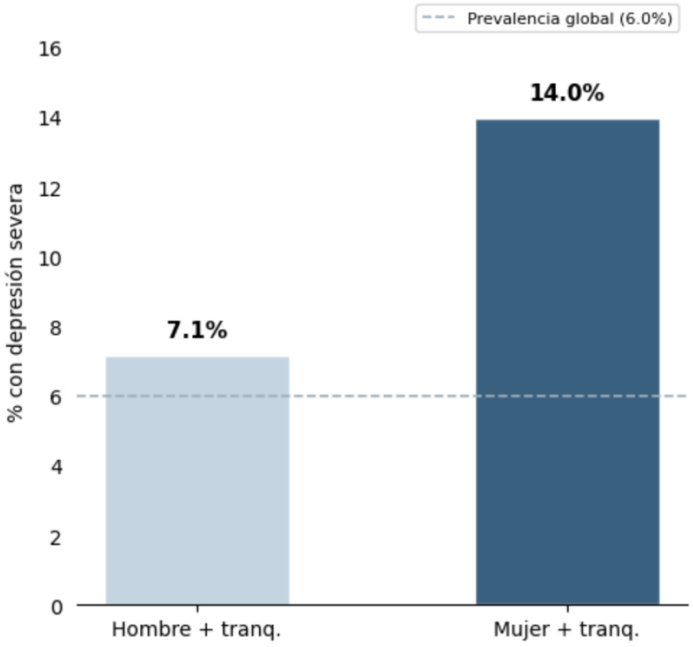
con lift superior a 3 y su posición como predictor individual más relevante, resultado coherente con el análisis SHAP del capítulo anterior.

Gráfico 21: Hombre sedentario vs Mujer sedentaria



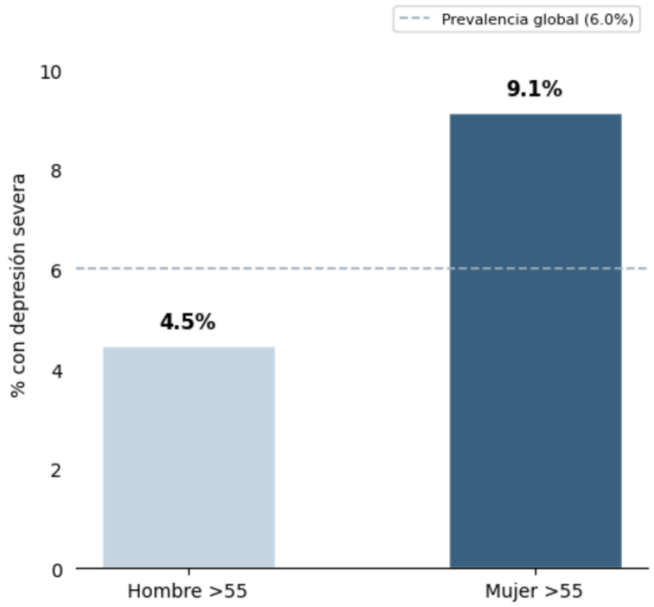
Fuente: Elaboración propia

Gráfico 22: Hombre + tranquilizantes vs Mujer + tranquilizantes



Fuente: Elaboración propia

Gráfico 23: Hombre mayor de 55 años vs Mujer mayor de 55 años



Fuente: Elaboración propia

Los Gráficos 21, 22 y 23 muestran el **efecto del sexo** dentro de distintos perfiles. En los tres casos, las **mujeres presentan una prevalencia claramente superior a la de los hombres con las mismas características**: 12,1 % frente a 5,9 % entre sedentarios, 14,0 % frente a 7,1 % entre consumidores de tranquilizantes, y 9,1 % frente a 4,5 % entre mayores de 55 años. Además, mientras que el hombre sedentario y el hombre mayor de 55 se sitúan por debajo de la prevalencia general, sus equivalentes femeninos la superan ampliamente. Esto confirma que el **sexo femenino** no es un factor de riesgo aislado, sino un **amplificador que eleva el** riesgo en cualquier perfil de vulnerabilidad, justificando su aparición recurrente en las reglas de mayor lift.

CAPÍTULO VII: CONCLUSIONES Y FUTURAS LÍNEAS DE INVESTIGACIÓN

1. CONCLUSIONES

El presente trabajo ha permitido analizar la depresión en la población adulta española desde una perspectiva multifactorial, utilizando una fuente de datos amplia y representativa como la Encuesta de Salud de España 2023 y aplicando técnicas de aprendizaje automático orientadas no solo a la predicción, sino también a la interpretación del fenómeno. El principal valor añadido del estudio reside en haber abordado la depresión como un problema complejo, condicionado por la interacción de variables sociodemográficas, clínicas y relacionadas con los estilos de vida, en lugar de simplificar su explicación a un único factor determinante. A continuación, se recogen las principales conclusiones obtenidas en cada fase del análisis.

Aportación metodológica

Desde una perspectiva metodológica, este trabajo realiza una aportación relevante. Por un lado, utiliza una **base de datos poblacional amplia y representativa**, lo que permite superar una limitación frecuente en la literatura, donde muchos estudios se basan en muestras muy específicas, como pacientes clínicos o estudiantes universitarios. Por otro lado, incorpora una **técnica de selección de variables** menos habitual en este ámbito, la información mutua, y la complementa con un análisis de interpretabilidad mediante **SHAP**, aportando una mayor transparencia y capacidad explicativa a los resultados obtenidos.

Análisis descriptivo

A nivel descriptivo, el análisis permitió **identificar varios patrones consistentes con la literatura previa**. La depresión fue más frecuente en personas de **mayor edad** (Ver Tabla 4, Gráfico 5 y Gráfico 6), en **mujeres** (Ver Tabla 6), en **individuos con menor nivel de actividad física** (Ver Tabla 5) y en determinadas Comunidades Autónomas (Ver Gráfico 10). En particular, el análisis territorial evidenció que la depresión no se distribuye de forma homogénea en España, sino que presenta una clara heterogeneidad territorial, con mayores niveles en comunidades como la **Comunidad Valenciana, Galicia o Madrid**, y menores en territorios como **Melilla, Extremadura o Ceuta**. Este resultado pone de manifiesto que el contexto territorial está relacionado con la presencia de depresión, lo que sugiere que las **estrategias de prevención e intervención en salud mental deberían**

adaptarse a las necesidades específicas de cada comunidad, con una asignación de recursos proporcional a su prevalencia, en lugar de diseñarse de forma uniforme a nivel nacional.

Asimismo, el análisis del **entorno afectivo y social** reveló que tanto el **apoyo social** percibido como la **convivencia en pareja** se asocian con una **menor prevalencia** de depresión (Ver Gráficos 7 y 8), mientras que la **pérdida del vínculo de pareja**, por viudedad, separación o divorcio, se asoció con las **tasas más elevadas** (Ver Gráfico 9). Estos hallazgos apuntan al aislamiento social como un factor de riesgo relevante y sugieren la conveniencia de incorporar en aten**ción primaria una valoración básica del contexto social del paciente** (si vive solo, si cuenta con apoyo cercano o si ha atravesado recientemente una situación de duelo o ruptura) con el fin de identificar de forma temprana situaciones de vulnerabilidad psicológica que, de otro modo, podrían pasar desapercibidas.

Selección de variables

En la fase de selección de variables mediante **información mutua**, ninguna variable alcanzó por sí sola un valor elevado (Ver Gráfico 11 y Tabla 9), lo que confirma que la **depresión no puede explicarse a partir de un único predictor** aislado. No obstante, las variables con mayor capacidad explicativa, y por tanto con mayor influencia en la presencia de depresión, fueron la **edad, la actividad física, la Comunidad Autónoma** de residencia y el **consumo de tranquilizantes**. Este hallazgo resulta coherente con la literatura científica previa, que ha identificado estos factores como elementos relevantes en la prevalencia y distribución de la depresión (Matute-Reyes y Torres-Palchisaca, 2025; Agudelo, Ante y Torres, 2017; Moreno, Lostao y Regidor, 2023).

Modelos predictivos e interpretabilidad

Desde una perspectiva predictiva, se desarrollaron y compararon dos modelos de clasificación: **regresión logística y XGBoost**. XGBoost obtuvo el mejor desempeño global (Ver Tabla 11 y Gráfico 12), mostrando una capacidad significativamente superior para identificar casos de depresión. Este resultado es coherente con la literatura previa, en la que **XGBoost suele destacar como uno de los algoritmos con mayor rendimiento** en tareas de clasificación similares. En esta línea, Lu, Wan y Liu (2025) observaron que

XGBoost y LightGBM fueron los modelos con mejor desempeño en la predicción de síntomas depresivos.

El análisis mediante **SHAP** sobre XGBoost reveló hallazgos especialmente relevantes desde una perspectiva preventiva. Las tres variables incluidas mostraron una **importancia global muy similar** (actividad física, edad y consumo de tranquilizantes), lo que indica que ningún factor domina de forma aislada, sino que los tres contribuyen de manera equilibrada a la probabilidad predicha de depresión (Ver Gráfico 14). En cuanto a la dirección de sus efectos (Ver Gráfico 15), niveles elevados de **actividad física** reducen la probabilidad predicha de depresión, actuando como **factor protector**, mientras que una **mayor edad y el consumo de tranquilizantes** la **incrementan**, configurándose como factores de riesgo. Estos resultados se encuentran en línea con la literatura previa, ya que Matute-Reyes y Torres-Palchisaca (2025), así como Agudelo, Ante y Torres (2017), han documentado una relación inversa entre actividad física y depresión. Del mismo modo, Moreno, Lostao y Regidor (2023) señalan que el riesgo de presentar problemas de salud mental aumenta con la edad.

Uno de los hallazgos más relevantes del estudio es que la **actividad física, un factor completamente modificable,** aparece como la variable con mayor peso dentro del modelo. Esto sugiere que el riesgo de depresión no está determinado únicamente por características biológicas o demográficas difícilmente alterables, sino que los hábitos de vida tienen una capacidad explicativa al menos igual de relevante. En términos prácticos, esto implica que **intervenciones orientadas a reducir el sedentarismo** podrían tener un **impacto real** y medible sobre la **prevalencia de la depresión** en la población adulta española. Entre las medidas más accesibles destacan la **prescripción de actividad física** desde atención primaria, el impulso de **programas municipales** de **deporte** accesibles o la **creación de entornos que favorezcan estilos de vida activos**, como espacios públicos adecuados o **campañas de concienciación** dirigidas a grupos de mayor riesgo.

Asimismo, el **consumo de tranquilizantes** emerge no solo como factor de riesgo, sino como un posible marcador de vulnerabilidad previa. Su presencia en el modelo sugiere que muchos individuos que consumen este tipo de medicamentos podrían estar atravesando o haber atravesado episodios depresivos no detectados o insuficientemente

tratados, lo que refuerza la **necesidad de un seguimiento más estrecho** por parte del **sistema sanitario** como mecanismo para favorecer la detección temprana.

Reglas de asociación

El análisis de reglas de asociación aportó una dimensión adicional al estudio, proporcionando aquello que los modelos predictivos no pueden ofrecer por sí solos: perfiles de riesgo concretos y directamente accionables (Ver Gráfico 19 y Tabla 13). El **perfil de máximo riesgo, mujer mayor de 55 años, sedentaria y consumidora de tranquilizantes**, presenta una prevalencia de depresión severa del 23,1 %, casi cuatro veces superior al 6,0 % de la población general, lo que lo convierte en un **criterio de detección directamente aplicable en atención primaria.**

De las cuatro reglas con mayor lift se extraen varias conclusiones prácticas. En primer lugar, el **consumo de tranquilizantes** es el **factor más consistente**: aparece en las cuatro reglas identificadas y, combinado con cualquier otro factor de riesgo, eleva sistemáticamente la prevalencia de depresión severa, lo que sugiere la necesidad de implementar **estrategias de seguimiento más activas para estos pacientes**.

Por otro lado, el **sexo actúa como amplificador transversal**: una **mujer** con el mismo perfil de riesgo que un hombre presenta sistemáticamente el doble de prevalencia de depresión severa, independientemente del factor considerado, ya sea ser sedentaria (12,1 % vs. 5,9 %), tomar tranquilizantes (14,0 % vs. 7,1 %) o tener más de 55 años (9,1 % vs. 4,5 %) (Ver Gráficos 21, 22 y 23), lo que justifica un **enfoque diferenciado por sexo en los protocolos de detección.** En conjunto, la coexistencia de varios factores multiplica el riesgo de padecer un cuadro de depresión severa, confirmando la naturaleza multifactorial y acumulativa de la depresión, idea central sobre la que se ha articulado este trabajo.

2. LIMITACIONES

El presente estudio presenta una serie de limitaciones que deben tenerse en cuenta al interpretar los resultados.

En primer lugar, la **fuente de datos** empleada presenta limitaciones inherentes a su naturaleza. La ESdE 2023, aunque es una encuesta robusta y representativa, **no está especializada en depresión**, sino que es una encuesta general de salud. Por ello, permite

aproximarse al fenómeno, pero no recoger con el mismo nivel de detalle variables clínicas más específicas sobre diagnóstico, tratamiento, comorbilidades o evolución de los síntomas. Además, **algunas variables** potencialmente relevantes, como el nivel de estrés laboral o la jornada laboral, tuvieron que **descartarse** debido a su elevado porcentaje de valores perdidos, lo que **limitó el análisis de ciertas dimensiones** del problema. A ello se suma que la **variable objetivo tiene carácter autoinformado**: la SEVERIDAD_DEPRESIVA refleja la **sintomatología percibida** por el propio encuestado y **no un diagnóstico clínico** realizado por un profesional sanitario, lo que puede introducir **sesgos de infradiagnóstico o sobrediagnóstico.**

En segundo lugar, el análisis se limita a los **datos de un único año** (2023). La incorporación de encuestas de **años anteriores** permitiría **analizar tendencias temporales** en la prevalencia de la depresión y en la influencia de los factores identificados.

Por último, el **rendimiento predictivo** de los modelos presenta **margen de mejora**. Los valores de AUC inferiores a 0,70 en ambos modelos son atribuibles en parte al número reducido de variables disponibles y a la complejidad intrínseca del fenómeno. Asimismo, los modelos **no fueron validados en un conjunto de datos externo** independiente, por lo que su capacidad de generalización a nuevas poblaciones no puede garantizarse.

3. FUTURAS LÍNEAS DE INVESTIGACIÓN

A partir de las limitaciones expuestas, se abren varias líneas de investigación futura. Sería especialmente valioso **replicar el análisis con una base de datos más especializada** en depresión o con información clínica de mayor detalle. Resultaría de interés desarrollar **estudios con un enfoque longitudinal**, para analizar no solo la presencia de depresión en un momento determinado, sino también su **evolución** a lo largo del tiempo. Asimismo, la **validación de los perfiles de riesgo identificados** mediante las reglas de asociación en **muestras clínicas independientes** permitiría evaluar su utilidad real en contextos de prevención e intervención en salud mental. Asimismo, investigaciones futuras podrían recurrir a **metodologías más avanzadas**, como la **Inteligencia Artificial guiada por hipótesis,** que permite incorporar el conocimiento clínico existente en la construcción del modelo (Gallagher, 2024). De este modo, se favorecería una interpretación más sólida de

los resultados y se reduciría el riesgo de identificar relaciones meramente estadísticas sin una base médica clara. Por último, la incorporación de un **mayor número de algoritmos** de aprendizaje automático y su **validación en conjuntos de datos externos** mejoraría la robustez y generalización de los resultados.

DECLARACIÓN DE USO DE HERRAMIENTAS DE INTELIGENCIA ARTIFICIAL GENERATIVA

Por la presente, yo, Carolina Rivera Marassa, estudiante de Doble Grado en Derecho y Business Analytics de la Universidad Pontificia Comillas al presentar mi Trabajo Fin de Grado titulado "Análisis de la depresión en España mediante técnicas de *Machine Learning*", declaro que he utilizado la herramienta de Inteligencia Artificial Generativa ChatGPT u otras similares de IAG de código sólo en el contexto de las actividades descritas a continuación:

1. **Brainstorming de ideas de investigación:** Utilizado para idear y esbozar posibles áreas de investigación.
2. **Referencias:** Usado conjuntamente con otras herramientas, como Science, para identificar referencias preliminares que luego he contrastado y validado.
3. **Interpretador de código:** Para realizar análisis de datos preliminares.
4. **Corrector de estilo literario y de lenguaje:** Para mejorar la calidad lingüística y estilística del texto.
5. **Sintetizador y divulgador de libros complicados:** Para resumir y comprender literatura compleja.

Afirmo que toda la información y contenido presentados en este trabajo son producto de mi investigación y esfuerzo individual, excepto donde se ha indicado lo contrario y se han dado los créditos correspondientes (he incluido las referencias adecuadas en el TFG y he explicitado para que se ha usado ChatGPT u otras herramientas similares). Soy consciente de las implicaciones académicas y éticas de presentar un trabajo no original y acepto las consecuencias de cualquier violación a esta declaración.

Fecha: 11/06/2026

Firma: Carolina Rivera Marassa

BIBLIOGRAFÍA

Agudelo, A., Ante Chaves, C. y Torres de G., Y. (2017). Factores personales y sociales asociados al trastorno de depresión mayor, Medellín (Colombia), 2012. *Revista CES Psicología*, 10(1), 21–34. https://doi.org/10.21615/cesp.10.1.2

Alonso, R. y Olivos, C. (2020). La relación entre la obesidad y estados depresivos. *Revista Médica Clínica Las Condes*, 31(2), 130–138. https://doi.org/10.1016/j.rmclc.2020.02.004

Al Masud, G. H., Shanto, R. I., Sakin, I. y Kabir, M. R. (2025). Effective depression detection and interpretation: Integrating machine learning, deep learning, language models, and explainable AI. *Array*, 25, 100375. https://www.sciencedirect.com/science/article/pii/S2590005625000025

Álvarez Vélez, M.ª I. (2023). Reparto competencial en materia sanitaria y las complicaciones surgidas durante el estado de alarma en España. *Revista Española de Derecho Constitucional*, 101–127. https://doi.org/10.18042/cepc/redc.128.04

Amat Rodrigo, J. (2018). Reglas de asociación y algoritmo Apriori con R. *Ciencia de Datos*. https://cienciadedatos.net/documentos/43_reglas_de_asociacion

Baba, A. y Bunji, K. (2023). Prediction of mental health problem using annual student health survey: Machine learning approach. *JMIR Mental Health*, 10, e42420. https://doi.org/10.2196/42420

Barañao, F. (s. f.). Mutual Information: una herramienta analítica para cuantificar relaciones no lineales. CFA Institute. https://higherlogicdownload.s3.amazonaws.com/CFAI/8795e2c6-63da-468e-b728-7557e0af30ce/UploadedImages/CFA_Mutual_Information_-_Francisco_Baranao__docx.pdf

Barnett, V. y Lewis, T. (1994). *Outliers in statistical data* (3.ª ed.). John Wiley & Sons.

Biilah, M. A., Raihan, M., Akter, T., Alvi, N., Bristy, N. J. y Rehana, H. (2021). Human depression prediction using association rule mining technique. *International Conference on Innovative Computing and Communications*, 1388, 223–237. https://doi.org/10.1007/978-981-16-2597-8_19

BioDatev. (2024). Odds ratio: La clave para medir el impacto de tus variables. https://biodatev.com/odds-ratio-calculo-e-interpretacion/

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

Choi, K. W., Chen, C.-Y., Stein, M. B. et al. (2019). Assessment of bidirectional relationships between physical activity and depression among adults: A 2-sample Mendelian randomization study. *JAMA Psychiatry*, 76(4), 399–408. https://doi.org/10.1001/jamapsychiatry.2018.4175

Data Science Python Blog. (2023). Mutual Info Classif — Scikit-Learn explained. *DataSciencePythonBlog*. https://datasciencepythonblog.net/mutual-info-classif-scikit-learn/

Espinosa-Zúñiga, J. J. (2020). Aplicación de algoritmos Random Forest y XGBoost en una base de solicitudes de tarjetas de crédito. *Ingeniería, Investigación y Tecnología*, 21(3). https://www.scielo.org.mx/scielo.php?script=sci_arttext&pid=S1405-77432020000300002

Esri. (s. f.). Cómo funciona el algoritmo XGBoost. *ArcGIS Pro Documentation.* https://pro.arcgis.com/es/pro-app/latest/tool-reference/geoai/how-xgboost-works.htm

Fiorillo, A., de Girolamo, G., Šimunović, I. F., Gureje, O., Isaac, M., Lloyd, C., Mari, J., Patel, V., Reif, A., Starostina, E. et al. (2023). The relationship between physical and mental health: An update from the WPA Working Group on managing comorbidity of mental and physical health. *World Psychiatry*, 22(1), 1–14.

https://onlinelibrary.wiley.com/doi/epdf/10.1002/wps.21055?getft_integrator=sciencedirect_contenthosting&src=getftr&utm_source=sciencedirect_contenthosting

Fluharty, M., Taylor, A. E., Grabski, M. y Munafò, M. R. (2017). The association of cigarette smoking with depression and anxiety: A systematic review. *Nicotine & Tobacco Research*, 19(1), 3–13. 10.1093/ntr/ntw140

Gallagher, C. (2024). Mayo researchers invented a new class of AI to improve cancer research and treatments. *Mayo Clinic News Network*. https://newsnetwork.mayoclinic.org/discussion/mayo-researchers-invented-a-new-class-of-ai-to-improve-cancer-research-and-treatments/

Garrido-Aguirre, J., Manzini, E. y Perera-Lluna, A. (2023). Aprendizaje automático y salud mental: de la promesa a la aplicación clínica. *Innovation*, 25–30. https://www.wemindcluster.com/wp-content/uploads/2022/05/Brains_02_vol-2_innovation.pdf

Google Developers. (s. f.). Clasificación: Exactitud, recuperación, precisión y métricas relacionadas. https://developers.google.com/machine-learning/crash-course/classification/accuracy-precision-recall?hl=es-419

Guillén, M. y Alonso, M. T. (2020). Modelos de regresión logística. *Fundació Universitat Oberta de Catalunya*. https://openaccess.uoc.edu/server/api/core/bitstreams/e830ec11-40f0-4aec-bc45-8bc10b123f7e/content

Harris, C. R. et al. (2020). Array programming with NumPy. Nature, 585(7825). https://doi.org/10.1038/s41586-020-2649-2

Hasan, M. J., Shifat, S. H., Matubber, J., Hossain, R., Rahman, M. A., Haque, B. M. T. y Hossen, M. J. (2026). An in-depth exploration of machine learning methods for mental health state detection: A systematic review and analysis. *Frontiers in Digital Health*, 7, Article 1724348. https://doi.org/10.3389/fdgth.2025.1724348

Henares Montiel, J., Ruiz-Pérez, I. y Sordo, L. (2019). Salud mental en España y diferencias por sexo y por comunidades autónomas. *Gaceta Sanitaria*. https://doi.org/10.1016/j.gaceta.2019.03.002

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, 9(3). https://doi.org/10.1109/MCSE.2007.55

IM Médico. (2025). En 2023 se alcanzó la cifra más alta de trastornos depresivos en España. *IM Médico Hospitalario*. https://www.immedicohospitalario.es/noticia/50373/en-2023-se-alcanzo-la-cifra-mas-alta-de-trastornos-depresivos-en-esp.html

Instituto Nacional de Estadística. (2025). Encuesta de Salud de España (ESdE) 2023. Nota de prensa. https://www.ine.es/dyngs/Prensa/ESdE2023.htm

Jiménez-Hernández, E., Mendoza-Chávez, Y., Moctezuma-Bautista, S., Vélez-Díaz, G. y Cano-Estrada, A. (2022). Relación entre el apoyo social percibido y la depresión en adultos mayores con asistencia en centro gerontológico. *Gerokomos*, 33(4), 230–233. https://scielo.isciii.es/pdf/geroko/v33n4/1134-928X-geroko-33-04-230.pdf

Lang, Q., Liu, X., He, Y., Lv, Q. y Xu, S. (2020). Association between working hours and anxiety/depression of medical staff during large-scale epidemic outbreak of COVID-19: A cross-sectional study. *Psychiatry Investigation*, 1167–1174. 10.30773/pi.2020.0229

Lu, C., Wan, S. y Liu, Z. (2025). Determinants of depressive symptoms in multinational middle-aged and older adults. *Digital Medicine*, 8, 501. https://pmc.ncbi.nlm.nih.gov/articles/PMC12321985/

Lundberg, S. y Lee, S. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30. https://doi.org/10.48550/arXiv.1705.07874

MathWorks. (s. f.). Overfitting: qué es y cómo evitarlo. *MATLAB & Simulink*. https://es.mathworks.com/discovery/overfitting.html

Matute-Reyes, C. F. y Torres-Palchisaca, Z. G. (2025). Niveles de depresión asociados a la actividad física en estudiantes de la carrera de Pedagogía de la Actividad Física y Deporte de la Universidad Católica de Cuenca, sede Azogues. *Resistances, Journal of the Philosophy of History*, 6(11). http://doi.org/10.46652/resistances.v6i11.213

Mayo Clinic. (2025). Depresión en las mujeres: información sobre la brecha entre los sexos. https://www.mayoclinic.org/es/diseases-conditions/depression/in-depth/depression/art-20047725

Mediavilla, D. (2024, 4 de noviembre). Estar soltero se relaciona con un riesgo de más del 80 % de tener depresión. *El País*. https://elpais.com/salud-y-bienestar/2024-11-04/estar-soltero-se-relaciona-con-un-riesgo-de-mas-del-80-de-tener-depresion.html

Ministerio de Asuntos Económicos y Transformación Digital. (2021). Guía práctica de introducción al análisis exploratorio de datos. https://datos.gob.es/sites/default/files/doc/file/analisis_exploratorio_de_datos_2021.pdf

Ministerio de Sanidad y Consumo. (2008). Prevención de los problemas derivados del alcohol. 1.ª Conferencia de prevención y promoción de la salud en la práctica clínica en España. https://www.sanidad.gob.es/areas/promocionPrevencion/jornadas/conferencia/docs/prevencionProblemasAlcohol.pdf;

Ministerio de Sanidad. (2023). Encuesta de Salud de España – ESdE. https://www.sanidad.gob.es/estadEstudios/estadisticas/encuestaSaludEspana/home.htm

Mitchell, R. y Frank, E. (2017). Accelerating the XGBoost algorithm using GPU computing. *PeerJ Computer Science*, 3. https://doi.org/10.7717/peerj-cs.127

Montesó-Curto, P. y Aguilar-Martín, C. (2013). Depresión según la edad y el género: análisis en una comunidad. *Atención Primaria*. 10.1016/j.aprim.2013.07.009

Moreno, A., Lostao, L. y Regidor, E. (2023). Características sociodemográficas y tendencia de los problemas de salud mental en España en las dos primeras décadas del siglo XXI. *Panorama Social*, (38). https://www.funcas.es/articulos/caracteristicas-sociodemograficas-y-tendencia-de-los-problemas-de-salud-mental-en-espana-en-las-dos-primeras-decadas-del-siglo-xxi/

Mouzo, J. (2022, 5 de julio). Cuánto pesa el estrés laboral en la salud mental: del 'burnout' a la depresión. *El País*. https://elpais.com/salud-y-bienestar/2022-07-05/cuanto-pesa-el-estres-laboral-en-tu-salud-mental-del-burnout-a-la-depresion.html

Mwiti, D. (2026). Algoritmo Apriori explicado: guía paso a paso con implementación en Python. *DataCamp*. https://www.datacamp.com/es/tutorial/apriori-algorithm

Na, K.-S., Cho, S.-E., Geem, Z. W. y Kim, Y.-K. (2025). Predicting future onset of depression among community dwelling adults in the Republic of Korea using a machine learning algorithm. *Neuroscience Letters*, 721, 134804. https://www.sciencedirect.com/science/article/abs/pii/S0304394020300744?via%3Dihub

Nair, M. (2023). Feature selection — Mutual information. *Medium*. https://medium.com/@miramnair/feature-selection-mutual-information-a0def943e1ed

O'Hagan, F. (2013). Mental health suffers among workers permanently impaired by job injury. *Canadian Journal of Public Health*. http://journal.cpha.ca/index.php/cjph/article/view/3036

Organización Mundial de la Salud. (2025). Trastorno depresivo (depresión). https://www.who.int/es/news-room/fact-sheets/detail/depression

Organización Mundial de la Salud. (2018, 21 de septiembre). El consumo nocivo de alcohol mata a más de 3 millones de personas al año, en su mayoría hombres. https://www.who.int/es/news/item/21-09-2018-harmful-use-of-alcohol-kills-more-than-3-million-people-each-year--most-of-them-men

Oviedo, J. (2023). MLZC25-19. Mutual Information Score: Selección inteligente de características para clasificación. *DEV Community*. https://dev.to/jesus_oviedoriquelme_084/mlzc25-19-mutual-information-score-seleccion-inteligente-de-caracteristicas-para-clasificacion-23b6

Pedregosa, F. et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830. https://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf

Pineda Sanjuán, S. (2025). Tratamiento de datos perdidos (Imputación múltiple con MICE). *GitBook*. https://silvia-pineda.gitbook.io/estudio-y-depuracion-de-datos/tema-3-valores-perdidos/tratamiento-de-datos-perdidos-imputacion-multiple-con-mice

Pousa Rodríguez, V., Miguelez Amboage, A., Hernández Blázquez, M., González Torres, M. Á. y Gaviria, M. (2015). Depresión y cáncer: una revisión orientada a la práctica clínica. *Revista Colombiana de Cancerología*, 19(3), 166–172. 10.1016/j.rccan.2015.04.005

Raschka, S. (2018). MLxtend: Providing machine learning and data science utilities and extensions to Python's scientific computing stack. *Journal of Open Source Software*, 3(24). https://doi.org/10.21105/joss.00638

Raschka, S. (2023). Association rules generation from frequent itemsets. mlxtend documentation. https://rasbt.github.io/mlxtend/user_guide/frequent_patterns/association_rules/

Remes, O., Mendes, J. F. y Templeton, P. (2021). Biological, psychological, and social determinants of depression: A review of recent literature. *Brain Sciences*, 11(12), art. 1633. https://pubmed.ncbi.nlm.nih.gov/34942936/

Retamal C., P. (1999). Depresión. Guías para el paciente y la familia. *Editorial Universitaria.* https://books.google.es/books?hl=es&lr=&id=1kwVmA7st_cC&oi=fnd&pg=PA9&dq=

la+depresion&ots=7RVjZvAV5U&sig=RJ83PQDdZjqNtguXD6rZRl4ZODA&redir_esc=y#v=onepage&q=la%20depresion&f=false

Rosales-Damián, G., Hidalgo-Rasmussen, C. A., Torres-Chávez, L. J. y Javier-Juárez, P. (2024). Relación entre consumo de riesgo de alcohol y síntomas depresivos en estudiantes universitarios de México. *Revista Internacional de Investigación en Adicciones*, 10(2). https://doi.org/10.28931/riiad.2024.2.05

Sahel, E. (s. f.). A dummy classifier: A baseline classifier or a null model. *Medium*. https://medium.com/@eskandar.sahel/a-dummy-classifier-a-baseline-classifier-or-a-null-model-71df50fd8947

Sánchez-Carro, Y., de la Torre-Luque, A., Leal-Leturia, I. et al. (2022). Importance of immunometabolic markers for the classification of patients with major depressive disorder using machine learning. *Progress in Neuro-Psychopharmacology & Biological Psychiatry*, 118, 110674. https://doi.org/10.1016/j.pnpbp.2022.110674

Scikit-learn Developers. (2024). Imputation of missing values. *Scikit-learn 1.5.0 User Guide*. https://scikit-learn.org/stable/modules/impute.html

Scikit-learn Developers. (s. f.). DummyClassifier. Scikit-learn. https://scikit-learn.org/stable/modules/generated/sklearn.dummy.DummyClassifier.html

Scikit-learn Developers. (s. f.). Métricas y puntuación: cuantificar la calidad de las predicciones. *Scikit-learn*. https://qu4nt.github.io/sklearn-doc-es/modules/model_evaluation.html

Sharma, N. y Verma, C. K. (2014). Association rule mining: An overview. *International Journal of Computer Science and Communication*, 5(1), 10–15. https://csjournals.com/IJCSC/PDF5-1/3.%20neesha.pdf

Shatte, A. B. R., Hutchinson, D. M. y Teague, S. J. (2019). Machine learning in mental health: A scoping review of methods and applications. *Psychological Medicine*, 49(9), 1426–1448. https://doi.org/10.1017/S0033291719000151

Simjanoski, M., Ballester, P. L., da Mota, J. C., de Boni, R. B., Balanzá-Martínez, V., Atienza-Carbonell, B., Bastos, F. I., Frey, B. N., Minuzzi, L., de Azevedo Cardoso, T. y Kapczinski, F. (2022). Lifestyle predictors of depression and anxiety during COVID-19: A machine learning approach. *Trends in Psychiatry and Psychotherapy*, 44. 10.47626/2237-6089-2021-0365

Suvendu Learns. (2024). Decoding Mutual Information (MI): A guide for machine learning practitioners. *Medium*. https://medium.com/@suvendulearns/decoding-mutual-information-mi-a-guide-for-machine-learning-practitioners-b0f0ca0b30c9

Tardivon, A. (2022). Random forest: Bosque aleatorio. Definición y funcionamiento. *Liora*. https://liora.io/es/random-forest-bosque-aleatorio-definicion-y-funcionamiento

The pandas development team. (2020). pandas-dev/pandas: Pandas [Software]. *Zenodo*. https://doi.org/10.5281/zenodo.3509134

Trelles, A., Fontaines Ruiz, T. y Ponce Rojo, A. (2025). Systematic review and meta-analysis of explainable machine learning models for clinical depression detection. *Behavioral Sciences*, 15(11), 1476. https://doi.org/10.3390/bs15111476

Virtanen, P. et al. (2020). SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nature Methods*, 17(3). https://doi.org/10.1038/s41592-019-0686-2

Walters-Williams, J. y Li, Y. (2009). Estimation of mutual information: A survey. *Lecture Notes in Computer Science*, 5816, 389–396. https://link.springer.com/chapter/10.1007/978-3-642-02962-2_49

Wang, C., Zhou, T., Fu, L., Xie, D., Qi, H. y Huang, Z. (2023). Risk and protective factors of depression in family and school domains for Chinese early adolescents: An association rule mining approach. *Behavioral Sciences*, 13(11), art. 893. https://doi.org/10.3390/bs13110893

Waskom, M. L. (2021). Seaborn: Statistical data visualization. *Journal of Open Source Software*, 6(60). https://doi.org/10.21105/joss.03021

Wei, Y., Qin, S., Liu, F., Liu, R., Zhou, Y., Chen, Y., Xiong, X., Zheng, W. y Zhang, R. (2025). Acoustic-based machine learning approaches for depression detection in Chinese university students. *Frontiers in Public Health*, 13, 1561332. https://doi.org/10.3389/fpubh.2025.1561332

Witten, I. H., Frank, E., Hall, M. A. y Pal, C. J. (2016). Data mining: Practical machine learning tools and techniques (4.ª ed.). p. 310. *Morgan Kaufmann*.

Xiong, X., Chen, J., Zhang, Y., Chen, X., Zhang, Y. y Ye, X. (2024). A quality control method based on physical constraints and data-driven collaborative for wind observations along high-speed railway lines. *EGUsphere*. https://www.researchgate.net/publication/381141119_A_quality_control_method_based_on_physical_constraints_and_data-driven_collaborative_for_wind_observations_along_high-speed_railway_lines

Zhai, X., Tong, H. H. Y., Lam, C. K. et al. (2024). Association and causal mediation between marital status and depression in seven countries. *Nature Human Behaviour*, 8, 2392–2405. https://doi.org/10.1038/s41562-024-02033-0

Zhou, Z.-H. (2021). Machine learning. *Springer Nature*. https://books.google.es/books?hl=es&lr=&id=ctM-EAAAQBAJ&oi=fnd&pg=PR6&dq=machine+learning+&ots=o_PhZ3Uw3u&sig=pIhcRLzXm2GlNFaDbb1JTamr8fo&redir_esc=y#v=onepage&q=machine%20learning&f=false

Zhu, A. (2021). Select features for machine learning model with mutual information. *Towards Data Science*. hhttps://towardsdatascience.com/select-features-for-machine-learning-model-with-mutual-information-534fe387d5c8/