

# Máster en Ciencia de Datos e Inteligencia Artificial

2025 – 2026

## Trabajo Grupal

**Título:**

*Diabetes Prediction*

**Estudiantes:**

Daniel Huarita, David Moya, Iñaki Asúa, Juan María Jiménez

## Resumen

---

El presente proyecto se desarrolló en el contexto del Máster en Data Science con el objetivo de realizar un análisis exploratorio de datos y sentar las bases para la construcción de modelos predictivos orientados a la detección de diabetes. El trabajo fue llevado a cabo por un equipo de cuatro integrantes bajo una metodología colaborativa basada en control de versiones con GitHub, permitiendo el desarrollo en paralelo y la integración sistemática de los avances en un notebook centralizado. El entorno de desarrollo se implementó en Python 3.11.9, utilizando bibliotecas especializadas para procesamiento y análisis de datos (*pandas*, *NumPy*, *OptBinning*), visualización (*Matplotlib* y *plotnine*) y análisis estadístico (*SciPy* y *StatsModels*), además de funciones auxiliares propias para estandarizar transformaciones y gráficos.

El análisis incluyó limpieza, transformación y discretización de variables, identificando como predictores más relevantes el índice de masa corporal (BMI), la hemoglobina glicosilada (HbA1c) y el nivel de glucosa en sangre, los cuales mostraron alta capacidad discriminante y relaciones monotónicas con la variable objetivo. Variables clínicas binarias como hipertensión y enfermedad cardíaca evidenciaron asociaciones significativas, mientras que factores demográficos y de estilo de vida aportaron información complementaria. A partir de estos hallazgos, el problema se formuló como una tarea de clasificación supervisada binaria, proponiéndose la evaluación comparativa de múltiples modelos, incluyendo regresión logística, árboles de decisión y métodos de ensamble, con el fin de seleccionar la alternativa que ofrezca el mejor equilibrio entre desempeño predictivo e interpretabilidad.

# CONTENIDOS

---

## ÍNDICE

<b>Resumen . . . . .</b>	<b>2</b>
<b>1 Primeras impresiones sobre los datos. . . . .</b>	<b>5</b>
1.1 Observaciones iniciales . . . . .	6
1.1.1 Variables categóricas . . . . .	6
1.1.2 Variables numéricas. . . . .	7
1.2 Valores duplicados. . . . .	9
1.3 Valores atípicos. . . . .	9
1.3.1 Edades no enteras . . . . .	10
1.3.2 IMC anormalmente altos . . . . .	10
1.4 Discretización de variables continuas . . . . .	11
<b>2 Relaciones entre diferentes variables . . . . .</b>	<b>13</b>
2.1 Variables continuas . . . . .	13
2.1.1 Edad y diabetes. . . . .	13
2.1.2 IMC y diabetes. . . . .	14
2.1.3 Edad e IMC segregados por diabetes. . . . .	15
2.1.4 Glucosa y HbA1c con la diabetes. . . . .	16
2.1.5 Edad y grupos de diferente historial de tabaquismo. . . . .	18
2.2 Variables categóricas o discretas. . . . .	19
2.2.1 Género. . . . .	19
2.2.2 Historial de tabaquismo. . . . .	20
2.2.3 Otras enfermedades. . . . .	21
2.2.4 Glucosa y Hb1Ac . . . . .	22
2.2.5 Edad e IMC . . . . .	23
<b>3 Test de Hipótesis. . . . .</b>	<b>25</b>
3.1 Chi-cuadrado ( $\chi^2$ ) y V de Cramer . . . . .	25
3.2 Test de homogeneidad de varianzas (Levene) y Kruskal-Wallis . . . . .	26

3.2.1	Kruskal-Wallis y Levene . . . . .	27
<b>4</b>	<b>Insights . . . . .</b>	<b>28</b>
<b>5</b>	<b>Metodología de trabajo y organización del equipo . . . . .</b>	<b>29</b>
5.1	Composición del equipo y organización . . . . .	29
5.2	Código, Entorno de desarrollos, librerías y otras herramientas. . . . .	29
5.2.1	Procesamiento y preparación de datos . . . . .	29
5.2.2	Visualización . . . . .	29
5.2.3	Análisis estadístico e inferencial . . . . .	30
5.2.4	Estructura del código y reutilización . . . . .	30
5.3	Flujo de trabajo general . . . . .	30
<b>6</b>	<b>Conclusiones y trabajo futuro . . . . .</b>	<b>31</b>
6.1	Trabajo futuro y modelado predictivo . . . . .	31
	<b>Referencias Bibliográficas . . . . .</b>	<b>33</b>

---

Documento generado el 8 de febrero de 2026

[jjimenez@afiglobaleducation.es](mailto:jjimenez@afiglobaleducation.es)

[iasua@afiglobaleducation.es](mailto:iasua@afiglobaleducation.es)

[shuarita@afiglobaleducation.es](mailto:shuarita@afiglobaleducation.es)

[dmoya@afiglobaleducation.es](mailto:dmoya@afiglobaleducation.es)

## 1 Primeras impresiones sobre los datos.

Este conjunto de datos, como hemos comentado en el resumen, está centrado en una serie de factores fisiológicos y demográficos relacionados con la salud. En concreto el conjunto de datos se centra en la diabetes. Para explicar la incidencia de diabetes en los diferentes individuos, vamos a comentar las variables disponibles acompañadas de una pequeña descripción en relación a su importancia en el diagnóstico de la diabetes:

**Edad - age:** La edad es un factor importante en la predicción del riesgo de diabetes. A medida que las personas envejecen, el riesgo de desarrollar diabetes aumenta. Esto se debe, en parte, a la reducción de la actividad física, a cambios hormonales y a una mayor probabilidad de desarrollar otras enfermedades crónicas relacionadas con la insulina.

**Género - gender:** El género puede influir en el riesgo de diabetes, aunque su efecto varía según el contexto. Por ejemplo, las mujeres con antecedentes de diabetes gestacional presentan un mayor riesgo de desarrollar diabetes tipo 2 en etapas posteriores de la vida.

**Índice de Masa Corporal - bmi:** El índice de masa corporal (IMC) es una medida de la grasa corporal basada en el peso y la altura del individuo. Un IMC elevado se asocia fuertemente con un mayor riesgo de diabetes tipo 2. El exceso de tejido adiposo, especialmente en la región abdominal, favorece la resistencia a la insulina y altera el metabolismo de la glucosa.

**Hipertensión - hypertension:** La hipertensión arterial es una condición que frecuentemente coexiste con la diabetes. Ambas enfermedades comparten factores de riesgo comunes y su relación es bidireccional. La presencia de hipertensión incrementa el riesgo de desarrollar diabetes tipo 2 y contribuye al deterioro de la salud cardiovascular.

**Enfermedad cardíaca - heart\_disease:** Las enfermedades cardiovasculares están asociadas con un mayor riesgo de diabetes. Esta relación es bidireccional, ya que ambas condiciones comparten factores de riesgo como la obesidad o la hipertensión.

**Historial de tabaquismo - smoking\_history:** El tabaquismo es un factor de riesgo modificable para la diabetes tipo 2. Fumar aumenta la resistencia a la insulina y afecta negativamente al metabolismo de la glucosa. La cesación del consumo de tabaco reduce de forma significativa el riesgo de desarrollar diabetes y sus complicaciones.

**Nivel de HbA1c - hba1c\_level:** La hemoglobina glucosilada (HbA1c) refleja el nivel medio de glucosa en sangre durante los últimos dos o tres meses. Valores elevados de HbA1c indican un mal control glucémico y se asocian con un mayor riesgo de diabetes y de complicaciones relacionadas con el corazón.

**Nivel de glucosa en sangre - blood\_glucose\_level:** La glucosa en sangre representa la concentración de glucosa circulante en un momento determinado. Niveles elevados,

especialmente en ayunas o tras la ingesta de carbohidratos, indican una alteración en la regulación glucémica y constituyen un criterio diagnóstico fundamental para la diabetes.

## 1.1 Observaciones iniciales

### 1.1.1 Variables categóricas

A primera vista, observamos **tres variables** que aparentan ser numéricas (*diabetes*, *enfermedad cardíaca* e *hipertensión*) realmente no lo son. En realidad, lo que tenemos es una codificación binaria donde 0  $\Rightarrow$  False y 1  $\Rightarrow$  True. Variables categóricas con codificación numérica. La proporción de individuos con cada una de estas enfermedades se resume en la tabla 1 a continuación:

Cuadro 1: Proporción de individuos con condiciones clínicas relevantes

Condición	Proporción (%)
Enfermedad cardíaca	3.94
Diabetes	8.50
Hipertensión	7.49

En cuanto al resto de variables categóricas, tenemos el género y el historial de tabaquismo del individuo. En primer lugar, para el género, hemos obtenido un total de tres categorías (*hombre*, *mujer* y *otros*). Al tener pocas observaciones para *otros* (solamente 18, lo que representa un 0,02% del total), decidimos **retirar dichas observaciones del estudio**.

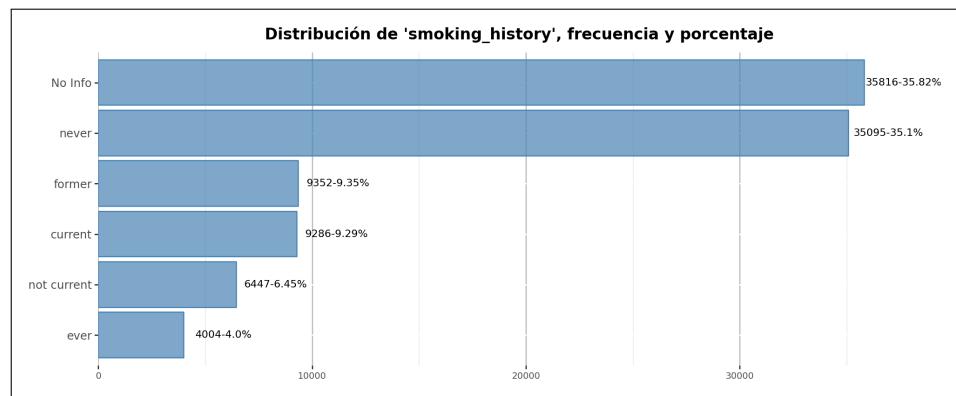


Figura 1: **Distribución del historial de fumadores**. Se muestra el conteo de individuos por categoría de historial de tabaquismo.

*Nota: La categoría No info es abundante y representa más del 35 % de los individuos.*

Del mismo modo, para la variable de historial de tabaquismo, tenemos una categoría **No info**, donde se incluyen todos aquellos individuos que no pertenecen al resto de las categorías. Sin embargo, en este caso, esta categoría es la más frecuente, con más de 35,000 observaciones, lo que representa un 35,82% del conjunto. Es evidente que, debido a la

cantidad de observaciones en esta categoría, no es apropiado deshacernos de ellas sin más. Por ello, mantenemos esta categoría como una propia.

### 1.1.2 Variables numéricas.

En nuestro conjunto de datos, las variables numéricas incluyen la *edad*, el *índice de masa corporal (IMC)*, el nivel de *HbA1c* y el nivel de *glucosa en sangre*.

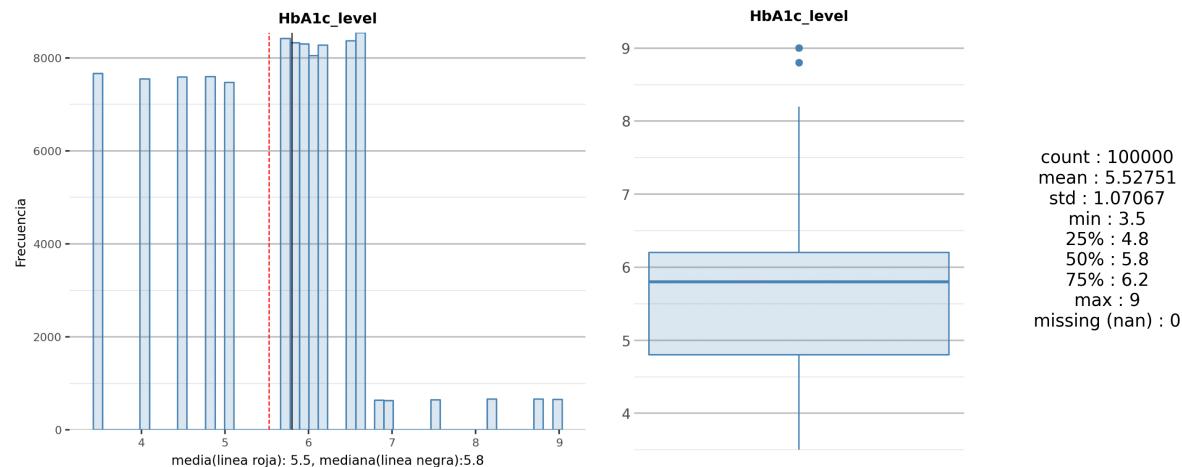
La variable edad presenta un rango aproximado de 0 a 80 años y una mediana de 43 años, lo que refleja una muestra heterogénea que incluye individuos jóvenes y de mayor edad, aspecto especialmente relevante considerando que el riesgo de diabetes tiende a incrementarse con la edad. Además será necesario razonar como afectan las observaciones de individuos jóvenes en el análisis. Hemos encontrado que existen unas dos mil observaciones donde la edad no es un número entero (aunque la variable sea continua, es raro que tengamos un valor de edad no entero, puesto que no se suelen medir así). Analizaremos estas observaciones en búsqueda de respuestas al porqué de estos valores y razonaremos su tratamiento.

El IMC también presenta una dispersión considerable, con valores que van desde alrededor de 10 hasta más de 70, con una mediana cercana de 27. Esto refleja la presencia de individuos con bajo peso, peso normal y sobrepeso u obesidad, lo que es coherente con la prevalencia de problemas metabólicos en la población general. Existen algunas observaciones pueden ser atípicas en la fisiología humana. Estas observaciones con IMCs mayores a 65 son escasas respecto al total de observaciones. Quizás en un futuro modelo estas observaciones podrían tener *high leverage* y modificar los resultados. Es por ello que eliminaremos estas observaciones.

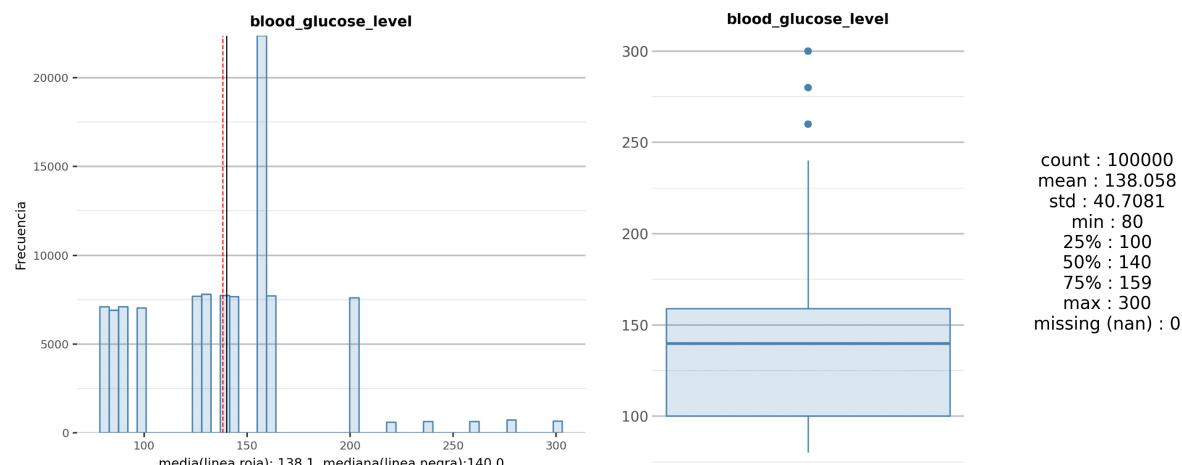
En cuanto al nivel de HbA1c, los valores oscilan entre 3,5 y 9, con un valor promedio de aproximadamente 5,8 %. La mayoría de los individuos se sitúa en rangos normales o levemente elevados, mientras que un pequeño grupo presenta niveles significativamente altos, lo que coincide con la proporción de personas con diabetes identificada previamente (8,5 %). Además, los valores de esta variable no son completamente continuos, sino que presentan diversos saltos. Por ejemplo, entre 5,1 y 5,5 no hay ninguna observación. Esto no es un problema en sí, pero abre la puerta a realizar una discretización de la variable.

Finalmente, el nivel de glucosa en sangre muestra valores entre 80 y más de 250 mg/dL, con una mediana alrededor de 140 mg/dL. La gran variabilidad de esta variable permite diferenciar claramente entre individuos con regulación normal de glucosa y aquellos con hiperglucemia o riesgo de diabetes. De nuevo, existen valores para los que no hay observaciones, por lo que sería interesante buscar una discretización de esta variable.

Estas dos posibles discretizaciones resultan interesantes, como veremos más adelante, por la grandísima relación (más que evidente) entre estas concentraciones de glucosa en sangre y la diabetes, y por cómo pueden ayudar a diferenciar individuos de alto y bajo riesgo. Además, es importante añadir que no hemos encontrado valores en ninguna de las dos variables que estén en contradicción con la realidad médica: no hay valores que correspondan a casos clínicamente imposibles.



(a) Distribución de HbA1c



(b) Distribución de glucosa

Figura 2: **Distribuciones de biomarcadores en la muestra.** Se muestran los histogramas de HbA1c (a) y glucosa (b) para facilitar la comparación visual de sus distribuciones.

*Nota: Los valores se muestran para la población total de estudio, sin separar por subgrupos.*

## 1.2 Valores duplicados.

---

En nuestro conjunto de datos no existe una **clave primaria** que identifique de manera única a cada observación. Sabiendo esto, nos preguntamos si existen filas idénticas. Dado que nuestras variables “continuas” realmente lo son a saltos (por ejemplo, la edad en enteros o la glucosa con mediciones discretas), esto no sería sorprendente. Sin embargo, al analizar los duplicados, observamos que el número de entradas que tienen al menos otra observación idéntica es de 6,939. En la tabla a continuación podemos revisar a detalle. [2](#).

Cuadro 2: Distribución de observaciones duplicadas exactas

Número de repeticiones	Número de patrones distintos
2	2523
3	431
4	85
5	29
6	9
7	5
8	1
9	2

*Nota: La tabla muestra el número de patrones de observaciones idénticas y su frecuencia de repetición en el conjunto de datos. Cada patrón corresponde a filas exactamente iguales en todas las variables, de estas tenemos 3085 únicas.*

A la luz de estos datos, no debemos preocuparnos en exceso por las variables repetidas por dos razones, en primer lugar el número de entradas repetidas no llega a representar un 4% del total del conjunto de datos. En segundo lugar, aquellas entradas repetidas no están repetidas demasiado, es decir, aquellas observaciones que se repiten lo hacen dos o tres veces aunque haya casos extremos como un par de observaciones que se repiten hasta nueve veces. Es por ello que no vamos a eliminar estas observaciones duplicadas, sino que las vamos a mantener en nuestro análisis. Como curiosidad vamos a mostrar un representante de cada uno de estos grupos con tantas repeticiones:

## 1.3 Valores atípicos.

---

Como hemos comentado en el apartado anterior, existen en concreto un par de variables cuyos valores son relativamente sospechosos. Estas variables eran las variables de **edad** y **IMC**. Vamos a analizar ambas por separado buscando explicar dichas observaciones y determinar el mejor rumbo a seguir en cuanto a su tratamiento.

Cuadro 3: Ejemplo de observaciones con perfil clínico idéntico y biomarcadores distintos

Sexo	Edad	Hipert.	Cardi	Tabaco	IMC	HbA1c	Glucosa	Diabetes	n_reps
F	80	0	0	No Info	27.32	6.5	159	0	9
F	80	0	0	No Info	27.32	6.2	90	0	9

Nota: Ambas observaciones comparten un perfil sociodemográfico y clínico idéntico y aparecen repetidas nueve veces en el conjunto de datos. Las diferencias se concentran únicamente en los biomarcadores continuos. Aquí vemos que los perfiles son similares, dos personas mayores, mujeres, sin información sobre su pasado fumador y que no presentan ninguna enfermedad.

Cuadro 4: Estadísticos descriptivos de edad e índice de masa corporal

Variable	Mean	Std	Min	P1	P25	P50	P75	P99	Max
Edad (años)	41.89	22.52	0.08	1.08	24.00	43.00	60.00	80.0	80.00
IMC	27.32	6.64	10.01	16.82	23.63	27.32	29.58	48.79	95.69

### 1.3.1 Edades no enteras

Una de las particularidades de este conjunto de datos es que, al describir las variables continuas, observamos que la edad presenta valores no enteros.

En concreto, el mínimo de edad es, según el cuadro 4, 0,08 años. El máximo es de 80 años, y los cuantiles parecen indicar que la distribución de la edad realmente llega hasta dichos valores. Pensando en un futuro modelo en el que podamos discretizar esta variable, con casi total seguridad estas edades tan cercanas pertenecerán al mismo grupo. Incluso si no lo hacemos, desde el punto de vista de la diabetes, no suele importar que un niño tenga 0,08 años o 0 años. Ante esta casuística, hemos decidido mantener las observaciones, pero **redondear el valor de la edad** para tener consistencia en cuanto al tipo de valores que conforman la variable `age`.

### 1.3.2 IMC anormalmente altos

Por otra parte, hemos observado (cuadro 4) que el valor máximo del IMC es superior a 95, mientras que el percentil 99 no llega a 50. Esto indica que estos valores de *IMC* no son simples observaciones situadas en la cola superior de una distribución, sino valores puntuales atípicos. Además, entre el mínimo y el percentil 99 se encuentran valores considerados relativamente normales en seres humanos, tal y como la propia *World Health Organization* indica en [este artículo](#).

Estos valores tan altos son potencialmente importantes a la hora de realizar un modelo en el sentido en el que estos pueden llegar a afectar significativamente el resultado del mismo. Tomando un corte en 65, tenemos únicamente una muestra de unos 47 individuos. Dichos individuos, tal y como se puede ver en el cuadro 5, presentan un mayor incidencia

de enfermedades como la diabetes y la hipertensión.

Cuadro 5: Estadísticos descriptivos y prevalencia de enfermedades por grupo de IMC.

Variable	Global	IMC $\geq 65$
<b>Enfermedades (proporción)</b>		
Diabetes	0.0850	0.3191
Enfermedad cardiovascular	0.0394	0.0638
Hipertensión	0.0749	0.1702
<b>Índice de masa corporal (kg/m<sup>2</sup>)</b>		
Media	27.32	72.92
Desviación típica	6.64	8.75
Percentil 99	48.79	95.47
Máximo	95.69	95.69

*Nota: El grupo IMC  $\geq 65$  corresponde a obesidad extrema severa. Valores de IMC superiores a este umbral se sitúan en la cola extrema de la distribución y presentan mayor prevalencia de enfermedades, además de alta influencia estadística en futuros modelos.*

Por todo lo anterior, se procede a excluir del análisis principal las observaciones con IMC  $\geq 65$ . Dichas observaciones corresponden a un subgrupo diferenciado, escasamente representado en la muestra y con un comportamiento estadístico claramente distinto.

## 1.4 Discretización de variables continuas

Una vez realizado este primer análisis, podemos plantear qué transformaciones previas aplicar a las variables. Observando la figura 2, se aprecian al menos dos zonas bien diferenciadas: una región con alta frecuencia de conteo, algo por debajo de 7, y otra de baja frecuencia por encima de ese valor, aproximadamente. Este patrón se repite también en la glucosa, aunque con un pico de frecuencias en torno a 150. En conjunto, estas distribuciones nos llevan a pensar que podemos dividir cada una de estas variables continuas en, al menos, dos conjuntos de valores: unos muy comunes y otros menos frecuentes.

Esta idea encaja con el concepto de enfermedad: cuando una persona enferma, los marcadores biológicos tienden a tomar valores poco comunes o fuera de la normalidad. Por todo ello, decidimos discretizar ambas variables y añadir también la edad a la lista de discretizaciones, para quedarnos exclusivamente con variables categóricas.

Para realizar la discretización utilizamos la librería `optbinning`, cuya función específica para este propósito, `OptimalBinning`, genera una partición óptima de una variable continua con respecto a una variable objetivo. En nuestro caso, diabetes. Esto implica que los puntos de corte se eligieron para maximizar el *information value* (IV) de la nueva variable

categórica.

Cuadro 6: Distribuciones discretizadas de Glucosa, HbA1c y Edad con IV.

Variable	Bin	Count	Count (%)	WoE	IV
HbA1c	(-∞, 5.75)	46262	46.27 %	1.7939	0.7448
	[5.75, 6.55)	41288	41.30 %	0.0946	0.0036
	[6.55, ∞)	12432	12.43 %	-1.8363	0.8314
<b>IV total</b>					<b>1.5798</b>
Glucosa	(-∞, 128.00)	35840	35.85 %	1.6376	0.5077
	[128.00, 159.50)	45558	45.57 %	0.1922	0.0156
	[159.50, 180.00)	7708	7.71 %	-0.0660	0.0003
	[180.00, ∞)	10876	10.88 %	-1.8042	0.6958
<b>IV total</b>					<b>1.2194</b>

Lo que podemos observar es que, al contrario de lo que habíamos observado a simple vista, se han creado de manera óptima 3 y 4 categorías para discretizar tanto la glucosa como el valor de HbA1c. La razón de esto estará en la relación que tenga la variable objetivo (diabetes) con ambas variables.

También decidimos segmentar las variables edad e IMC. Dado que son continuas y que uno de los modelos a considerar será la regresión logística.

Cabe remarcar que esta discretización es meramente exploratoria y que, al momento de construir los modelos predictivos, se debe realizar una partición en conjunto de entrenamiento y de prueba, discretizando únicamente sobre el de entrenamiento y trasladando ese mismo criterio al de prueba.

## 2 Relaciones entre diferentes variables

A continuación, nos adentramos de lleno en el análisis de los datos. No solo estudiaremos cómo son nuestras variables, sino también qué tipos de relaciones (si las hay) existen entre ellas. Para ello, dividimos el análisis en dos bloques principales.

El primero corresponde al apartado 2.1, donde tomaremos las variables continuas (antes de haber realizado el *binning*) y observaremos sus relaciones con otras variables, tanto categóricas (mediante la segregación de observaciones) como continuas, a través de gráficos de dispersión.

En un segundo bloque (apartado 2.2), analizaremos las relaciones entre las variables categóricas, prestando especial atención a la variable *diabetes*. Esto incluirá las nuevas variables discretizadas de manera óptima en el apartado 1.4.

### 2.1 Variables continuas

En esta sección del análisis nos hemos apoyado principalmente en las siguientes técnicas de representación. Histogramas superpuestos en función de la variable objetivo *diabetes* y de boxplot.

Para complementar el análisis descriptivo y evitar conclusiones basadas únicamente en la inspección visual, se recurrió a la aplicación de dos contrastes de hipótesis.

- **Test de Levene:** Hipótesis nula, igualdad de varianzas entre poblaciones.
- **Kruskal-Wallis:** Hipótesis nula, todas las poblaciones/grupos tienen la misma distribución.

Describimos algunos de estos tests y sus estadísticos en la sección 3.

#### 2.1.1 Edad y diabetes.

Representamos la variable numérica *edad* segregada por la variable categórica correspondiente al diagnóstico de diabetes. A simple vista, se observa que ambos grupos presentan distribuciones muy diferentes. El *boxplot* y el histograma muestran que, en los individuos sanos, la distribución de la edad es bastante uniforme, aunque decrece ligeramente en el rango superior.

No obstante, para los individuos con diabetes, la distribución de la edad es claramente distinta: **hay pocas observaciones en los rangos inferiores de edad**, que además aparecen como *outliers*. Sin embargo, a medida que la edad aumenta, la **densidad** de personas con diabetes va aumentando. Es decir, las personas con diabetes no se distribuyen de manera uniforme a lo largo de todos los rangos de edad, sino que se concentran en la parte alta del intervalo.

Por último, también se observa una diferencia clara en las medias: el grupo no diabético tiene una media de edad de  $\sim 40$  años, mientras que el grupo diabético se sitúa, en promedio, en torno a  $\sim 60$  años.

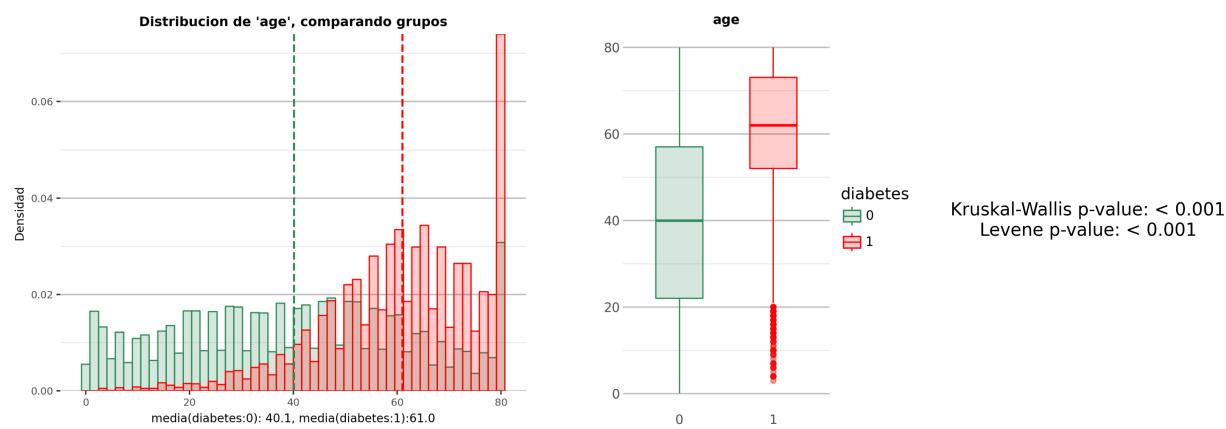


Figura 3: **Edad versus diabetes.** Distribución de la edad segregada por diagnóstico de diabetes.

Tanto el test de Levene de homocedasticidad, como el test de Kruskal-Wallis se rechazan de manera que podemos verificar que ambas poblaciones no tienen una distribución similar ni varianzas similares tampoco. De manera visual esto no es una sorpresa pues es evidente que la distribución de edades es más o menos uniforme en los no diabéticos, pero no es así en los diabéticos. En resumen, las personas diabéticas suelen desarrollarla a edades tardías, de manera que aunque una persona tenga un mal estilo de vida, tendrá mas riesgo de tener diabetes sí, pero quizás no la desarrolle hasta una edad más tardía.

## 2.1.2 IMC y diabetes.

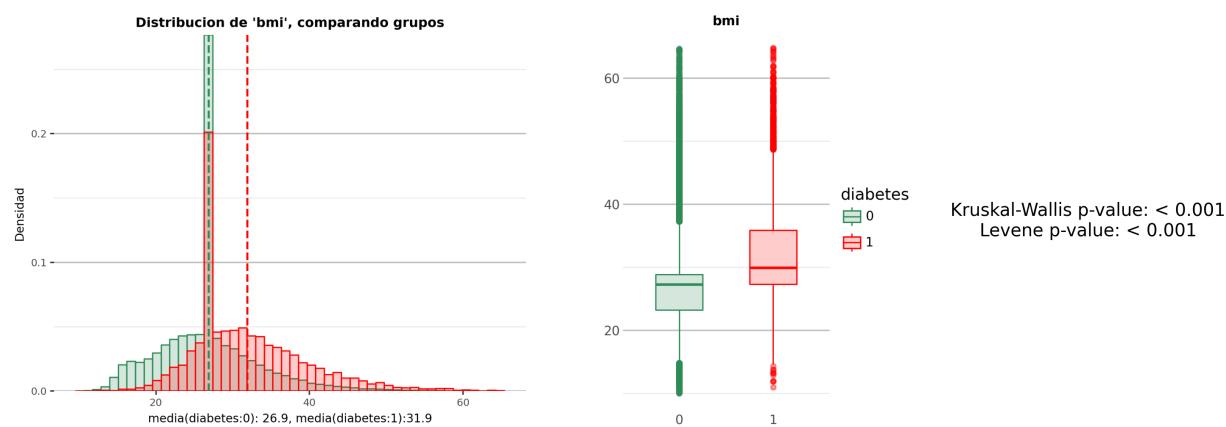


Figura 4: **IMC versus diabetes.** Distribución del índice de masa corporal segregado por diagnóstico de diabetes.

Observamos dos cosas. En primer lugar, al contrario que en el caso anterior, se aprecia una gran similitud entre las distribuciones. Ambas tienen una forma de campana y presentan una “explosión” de densidad en torno al valor 27 de IMC, que puede considerarse un valor promedio y que coincide con la media de las observaciones no diabéticas. Sin embargo, en el grupo no diabético la campana alcanza su máxima densidad en torno a ese valor, mientras que en los diabéticos se extiende algo más hacia valores altos de IMC y presenta una cola más pesada, que llega más lejos que la de los no diabéticos.

En segundo lugar, se rechazan ambos tests de hipótesis. Aunque los dos histogramas son relativamente similares, es posible que no comparten la misma mediana (en el *boxplot* se observa precisamente una mediana mayor en los diabéticos), lo que concuerda con el rechazo del test de Kruskal-Wallis bajo la observación de formas similares en ambas distribuciones. También se rechaza la homogeneidad de varianzas, lo cual no es de extrañar dada la mayor extensión de valores que cubren los diabéticos respecto a los sanos.

En resumen, la población no diabética está más concentrada alrededor de su media de IMC, en torno a 27 (valor relativamente saludable), mientras que los diabéticos se distribuyen de manera más extendida hacia valores más altos, con media y mediana superiores al grupo sano. Esto pone de manifiesto la estrecha relación entre el IMC y la diabetes: un mayor IMC implica, en general, una peor salud y una mayor probabilidad de diabetes.

### 2.1.3 Edad e IMC segregados por diabetes.

En este caso, analizamos dos variables continuas, que acabamos de describir en función de la diabetes. Nuestro objetivo es ver cómo se distribuyen las personas diabéticas en función de su edad y su IMC. Como ya hemos visto, las personas diabéticas presentan distribuciones de edad e IMC desplazadas hacia valores más altos que sus contrapartes sanas. Esto puede observarse en la siguiente figura:

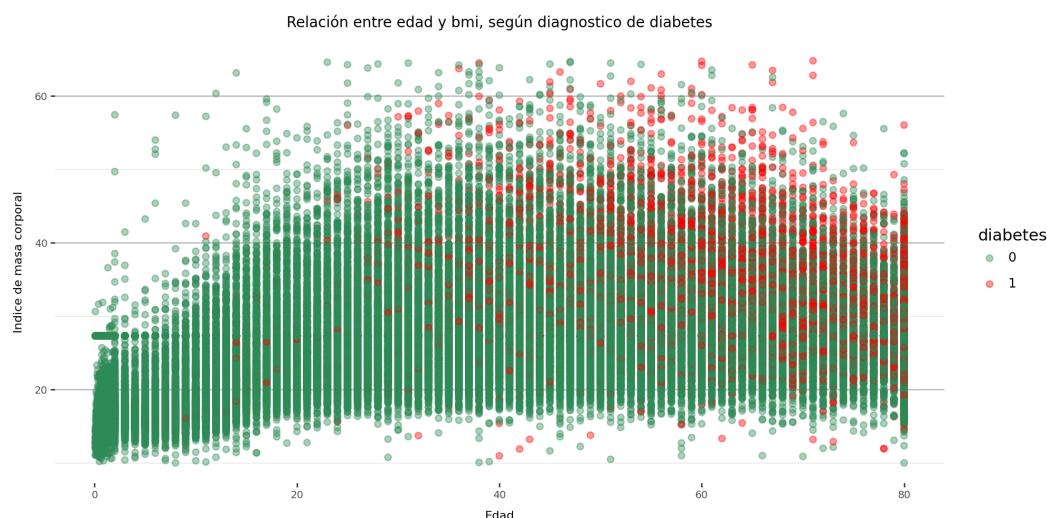


Figura 5: **Edad e IMC segregados por diabetes.** Distribución conjunta de edad e IMC para individuos con y sin diagnóstico de diabetes.

Se puede observar que, a medida que aumenta la edad, la concentración de puntos rojos (diabéticos) crece incluso para valores de IMC relativamente normales. No obstante, para valores de edad más bajos, los puntos rojos aparecen en valores de IMC superiores a la normalidad ( $\sim 27$ ). Esto sugiere que la edad es un factor importante a la hora de determinar si una persona tiene o no diabetes, pero también lo es la interacción entre ambas variables. En particular, una persona con un IMC alto tenderá a desarrollar diabetes a una edad más temprana que otra con un IMC habitual.

#### 2.1.4 Glucosa y HbA1c con la diabetes.

Estas dos variables están relacionadas de manera lógica. Resulta intuitivo que, cuanto mayores sean los niveles de glucosa en los individuos, mayores serán sus probabilidades de ser diagnosticados de diabetes. De hecho, estas variables juegan un papel fundamental en el diagnóstico de la enfermedad, como se puede observar en la figura 6:

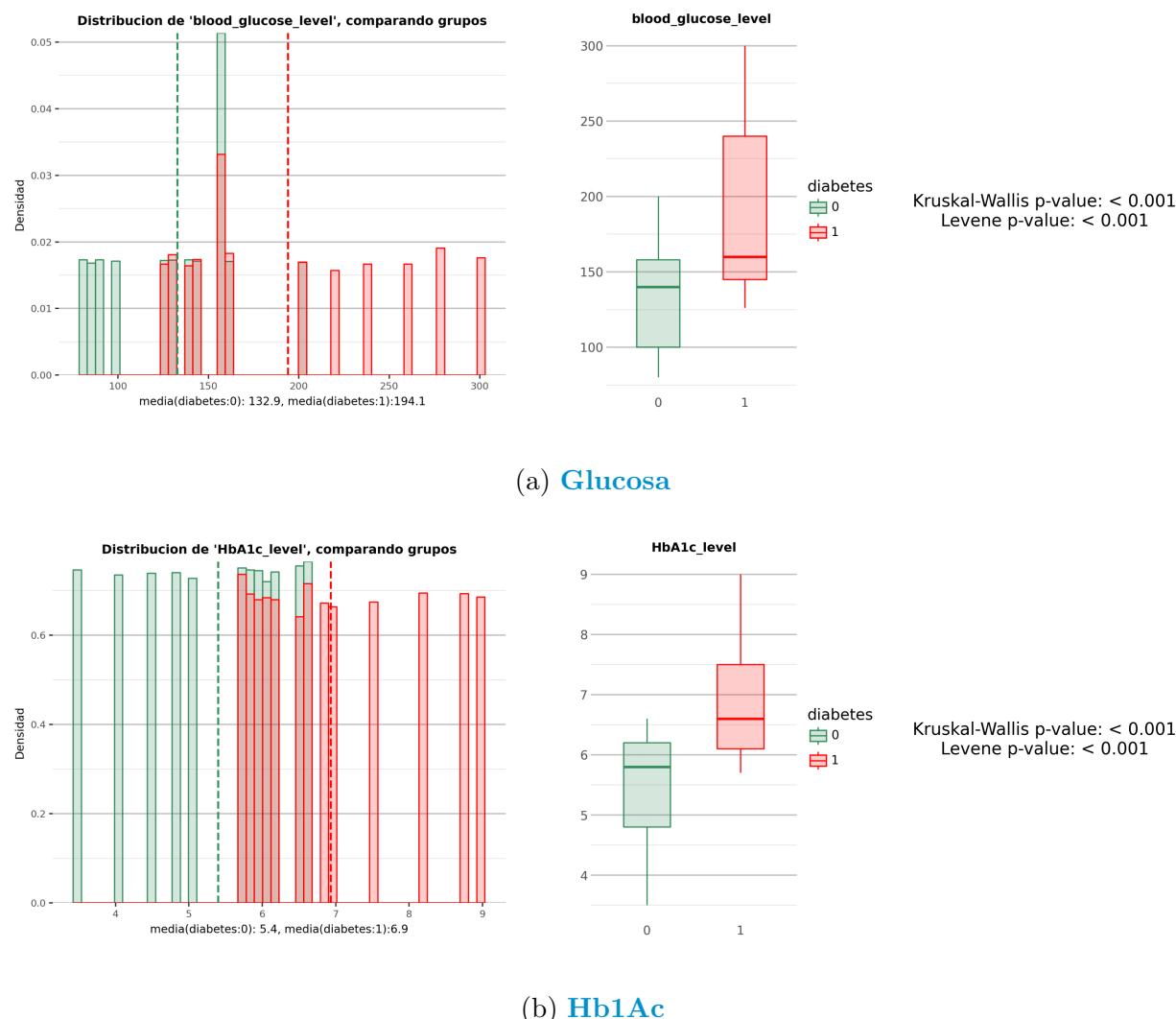


Figura 6: **Glucosa y HbA1c versus diabetes.** Distribuciones segregadas por diagnóstico de diabetes.

Lo que podemos observar con bastante claridad es que ambas variables están distribuidas de manera relativamente uniforme entre dos valores: un mínimo y un máximo. Lo interesante aquí es que los rangos de valores de ambas poblaciones apenas solapan, y lo hacen en una región muy pequeña. Esto indica que estas variables separan muy bien a las poblaciones con distintos diagnósticos de diabetes.

Según [esta referencia](#), los valores típicos para una persona sin diabetes se sitúan **entre 70 y 180 mg/dL en glucosa** y **entre 4 y 6 % en HbA1c**. Estos valores se ajustan muy bien a nuestras observaciones.

Las zonas de solapamiento pueden corresponder a individuos con uno de los dos valores en rango normal y el otro en rango diabético, cuyo diagnóstico, por tanto, es más límite. Es evidente que estas variables definen dos poblaciones con medias y medianas distintas. Esto se puede corroborar, para la variable de glucosa, mediante el test de Kruskal-Wallis, teniendo en cuenta que las distribuciones son similares en ambos grupos. Para la variable HbA1c no es tan directo, pero la observación de los *boxplots* también sugiere esta diferencia de medianas.

En definitiva, ambas variables son esenciales a la hora de determinar si una persona es o no diabética. Esto no solo se aprecia en los datos, sino que, además, el conocimiento médico apoya esta observación.



Figura 7: **Glucosa y HbA1c segregadas por diabetes.** Diagrama de dispersión de ambas variables separando en gran medida a individuos diabéticos y no diabéticos.

*Nota: Como vemos, en cada variable existe una pequeña región de solapamiento entre individuos con y sin diabetes. Sin embargo, fuera de esa región se generan cuatro zonas en las que los individuos son, o bien todos diabéticos, o bien todos sanos.*

## 2.1.5 Edad y grupos de diferente historial de tabaquismo.

Para finalizar este análisis de las variables continuas, analizamos qué ocurre con los distintos grupos según su historial de tabaquismo. Como habíamos comentado anteriormente, nos encontramos ante una serie de categorías en las que, en algunos casos, no sabemos exactamente qué diferencias hay entre ellas. Dado que la edad es un factor importante a la hora de diagnosticar la diabetes, analizaremos si los distintos grupos de tabaquismo presentan distribuciones de edad diferentes.

Cuadro 7: Edad media por historial fumador

Estadístico	No Info	never	current	not current	ever	former
Media (años)	33.54	43.89	44.07	47.72	49.14	57.06

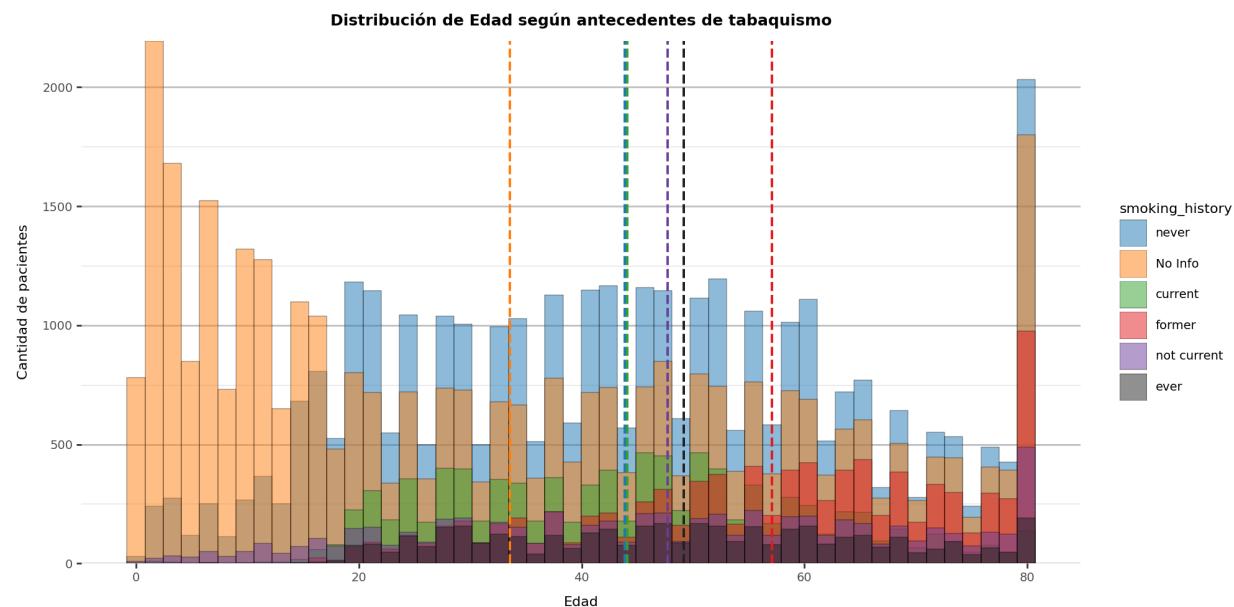


Figura 8: **Edad por historial fumador.** Distribución de la edad en función de las categorías de historial fumador.

Lo que podemos ver es lo siguiente:

- **No info:** Parece referirse, en gran medida, a niños y menores en las zonas bajas del espectro. Después, el conteo se distribuye de manera uniforme hasta los rangos de edad más altos, donde empieza a disminuir. La gran cantidad de menores y niños que hay en este grupo podría indicar una menor incidencia futura de diabetes, debido a la corta edad de una parte importante de sus integrantes.
- **Never:** Este grupo de no fumadores está distribuido de manera bastante uniforme entre los 20 y los 60 años. En valores fuera de este intervalo presenta poca concentración. Quizás la gente mayor de 60 años tenga menos probabilidades de no haber fumado nunca, y los menores de 20 quizás se registren como **No info**.

- **Current:** Este grupo es menos abundante que los anteriores, como ya se comentó. Además, se concentra en edades entre los 20 y los 60 años, con una distribución relativamente uniforme.
- **Ever y Not current:** Tienen distribuciones muy similares. Ambas se concentran mayormente entre los 20 y los 80 con un ligero pico en torno a los 50. No obstante la variable **not current** tiene una ligera cola en la zona inferior de edades, por debajo de los 20 años, lo cual genera que su media sea algo inferior a la del grupo **ever**.
- **Former:** Este grupo es sin duda uno de los más interesantes. Su distribución se concentra sobretodo en individuos por encima de los 35 – 40 años hasta los 80 años. Aquí se junta una población relativamente envejecida con un hábito nocivo para la salud como lo es fumar. La interacción de estos dos sucesos podría llevar a observaciones de incidencia de diabetes más altas de lo normal en este grupo.

## 2.2 Variables categóricas o discretas.

---

Para analizar las variables categóricas y las relaciones existentes entre ellas, nos hemos basado en el estudio de histogramas de conteo. Hemos separado cada categoría por la clase de diabetes (u otras enfermedades) para observar la incidencia de estas condiciones en cada clase de las variables categóricas y detectar si alguna población es más propensa a desarrollarlas.

Este análisis se complementa con la realización de los siguientes test.

- Chi-cuadrado ( $\chi^2$ ): Hipótesis nula, las dos variables categóricas son independientes.
- V de Cramér: Mide la fuerza de asociación entre variables categóricas, con un rango entre (0,1), donde 0 indica que no hay asociación y 1 asociación más fuerte.

### 2.2.1 Género.

En primer lugar, detallamos brevemente la relación (o, mejor dicho, la ausencia de relación) que encontramos entre la variable género y la diabetes. Al segregar entre diabéticos y no diabéticos para hombres y mujeres, no se observan grandes diferencias:

Cuadro 8: Porcentaje de diabetes por género

Género	No diabetes (%)	Diabetes (%)
Female	92.40	7.60
Male	90.25	9.75
Total	91.51	8.49

No obstante, hay que matizar que el test  $\chi^2$  resulta significativo según el valor  $p$  obtenido. Esto, sin ninguna duda, se debe a la gran cantidad de observaciones disponibles: ante tamaños muestrales grandes, los estadísticos  $\chi^2$  pueden volverse significativos incluso ante

pequeñas diferencias. Sin embargo, el estadístico  $V$  de Cramér indica que, si bien parece existir una asociación entre género y diabetes, su fuerza es pequeña ( $V = 0,038$ ).

Por tanto, podemos concluir que, de existir una relación entre género y diabetes, su magnitud sería reducida (como sugiere la tabla) y que los hombres serían ligeramente más propensos a padecer diabetes que las mujeres.

### 2.2.2 Historial de tabaquismo.

De la misma forma que para la variable anterior, hemos buscado observar una relación de dependencia entre las diferentes categorías de historial de tabaquismo y la incidencia de diabetes. Lo que podemos ver es que existe una gran diferencia entre las categorías, *Never*, *Ever*, *Not current* con las *No info*, *Former*.

Como ya habíamos previsto, la longevidad de la categoría *Former* y, por otro lado, la relativa juventud de *No info* determinan en gran parte la incidencia de la diabetes, al presentar la mayor y la menor incidencia, respectivamente. Sin embargo, las categorías de edad intermedia muestran valores de incidencia de diabetes relativamente similares.

Cuadro 9: Porcentaje de diabetes según historial de tabaquismo

Historial de tabaquismo	No diabetes (%)	Diabetes (%)
No Info	95.94	4.06
never	90.49	9.51
current	89.81	10.19
not current	89.28	10.72
ever	88.21	11.79
former	83.00	17.00
Total	91.51	8.49

Esto no quiere decir que el historial de tabaquismo no sea importante en el diagnóstico de diabetes, sino que quizás tenga un efecto menor que el de la edad o que su efecto sea interactivo: el hecho de haber fumado importa más cuanto mayor es la persona. En cualquier caso, existen indicios significativos de relación entre estas categorías y la incidencia de diabetes, y la asociación es relativamente intensa ( $V = 0,14$ ).

En cualquier caso, hemos decidido realizar un último análisis sobre cómo afecta cada una de estas categorías al diagnóstico de otras enfermedades en la figura 9

Lo que podemos observar es que las relaciones entre las diferentes categorías y las enfermedades siguen un patrón bastante similar. La categoría *former* se mantiene como la de mayor incidencia de enfermedades (y también la más longeva), mientras que *no info* es la de menor incidencia (y la más joven). El resto de categorías suelen presentar niveles de incidencia bastante similares, moviéndose, como mucho, en 3 puntos porcentuales entre sí. La excepción se da en el caso de la enfermedad cardiaca: las categorías *current*, *never* y *not current* sí mantienen esa similitud en la incidencia, pero la categoría *ever* presenta una

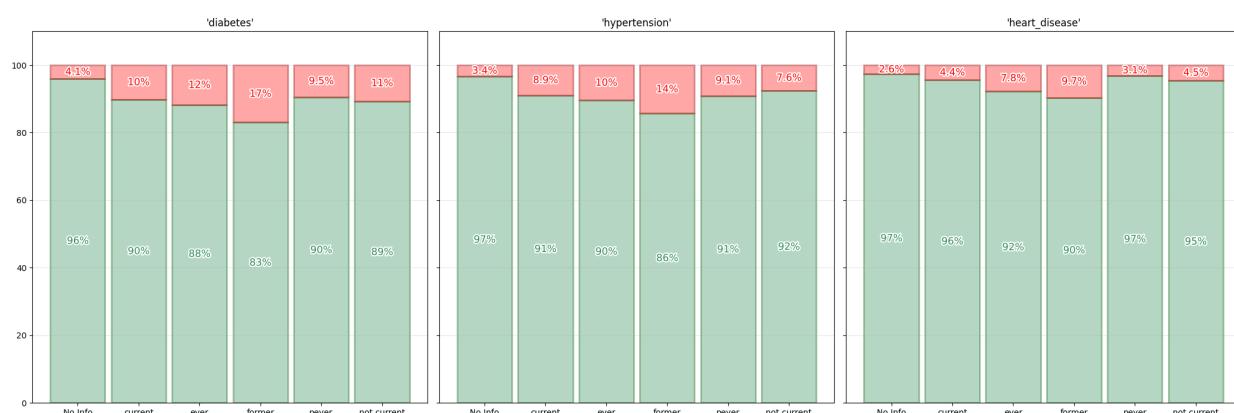
(a) **Histórico de tabaquismo y enfermedad**

Figura 9: **Relación entre histórico de tabaquismo y enfermedades.** Incidencia de cada enfermedad según el histórico de tabaquismo.

incidencia algo mayor y más cercana a la de *former*.

### 2.2.3 Otras enfermedades.

Para visualizar mejor la relación entre distintas enfermedades, hemos tomado la diabetes como referencia. En función de si los individuos son diabéticos o no diabéticos, analizamos qué porcentaje de la población desarrolla alguna otra enfermedad. Si la coexistencia de enfermedades es más común que el padecimiento aislado de una de ellas, podremos concluir que existen relaciones de riesgo que sugieren que una puede derivar en otra o que los hábitos que causan una también pueden ser factores de riesgo para otra.

Ante estas observaciones tenemos los siguientes resultados:

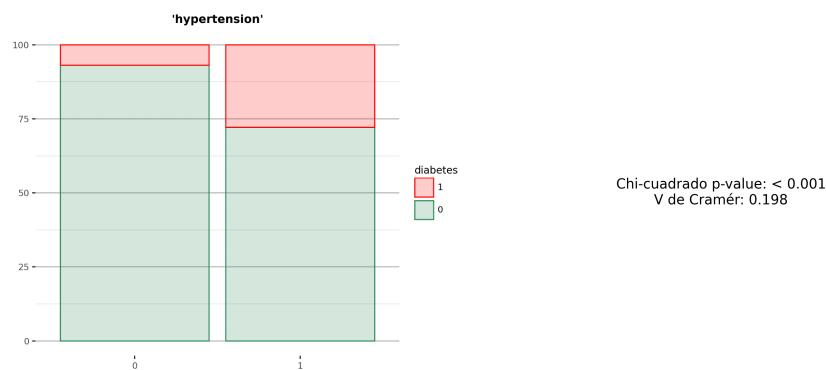
Cuadro 10: Porcentaje de diabetes según hipertensión y enfermedad cardíaca

<b>Variable</b>	<b>Categoría</b>	<b>No diabetes (%)</b>	<b>Diabetes (%)</b>
Hypertension	No	93.08	6.92
	Sí	72.11	27.89
Heart disease	No	92.48	7.52
	Sí	67.83	32.17

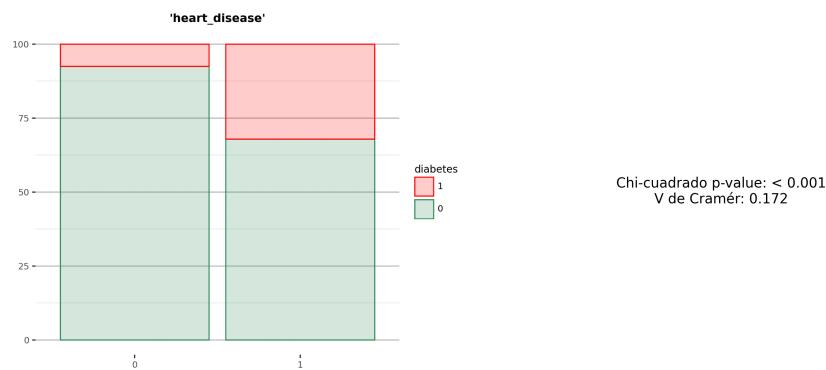
El patrón observado es claro. En el caso de la hipertensión, la proporción de individuos con diabetes es aproximadamente cuatro veces mayor entre quienes presentan esta condición (27,89 %) en comparación con quienes no la padecen (6,92 %). Un comportamiento muy similar se observa en la enfermedad cardíaca, donde la prevalencia de diabetes alcanza el 32 % frente al 7,52 % en individuos sin dicha patología.

Estos resultados evidencian una asociación fuerte entre la diabetes y las enfermedades cardiovasculares analizadas, lo que queda respaldado por el test de independencia  $\chi^2$ , cuyo valor p es inferior a 0,001 en ambos casos. No obstante, dado el gran tamaño muestral,

este resultado no resulta sorprendente y debe interpretarse con cautela. Para evaluar la magnitud real de la asociación, se ha considerado el estadístico V de Cramér, que toma valores de  $V \geq 0,17$  en ambas relaciones. Este nivel indica una asociación de intensidad moderada-alta.



(a) Hipertensión



(b) Enfermedad cardíaca

Figura 10: **Comorbilidades cardiovasculares versus diabetes.** Distribución del diagnóstico de diabetes en función de la presencia de hipertensión y enfermedad cardíaca.

Por último, es importante destacar que, a partir de estos análisis, no puede establecerse una relación de causalidad entre las enfermedades. Los resultados únicamente permiten concluir la existencia de una asociación significativa, por lo que sería necesario otro tipo de análisis para determinar posibles relaciones de dependencia directa o causal.

#### 2.2.4 Glucosa y Hb1Ac

Como ya comentamos anteriormente en el análisis continuo de la sección 2.1, estas dos variables están estrechamente ligadas al diagnóstico de diabetes. Ahora vamos a utilizar la discretización de la sección 1.4 para determinar si las diferentes categorías que hemos creado en cada variable guardan una relación directa con la diabetes, tal y como tratábamos de evaluar mediante el *optimal binning*. Para ello, compararemos la discretización por cajas predeterminadas (histograma) con nuestra discretización óptima.

En la figura 11 uno de los resultados más destacables es la monotonía. Nuestras clases muestran que, al pasar de una clase con niveles más bajos de glucosa a otra con niveles más altos, se incrementa la incidencia de diabetes.

Si nos remitimos a la tabla 6, podemos ver que las clases extremas son precisamente las que presentan mayores valores de *IV*, mientras que este *information value* es mucho más pequeño en las clases centrales. La razón es que, al separar la población por la variable diabetes, las categorías centrales (en las que encontramos individuos con valores frontera entre ser o no diabéticos) generan mayor confusión y, por tanto, no separan tan bien a la población.



Figura 11: **Distribución de diabetes** según niveles de glucosa y HbA1c.

## 2.2.5 Edad e IMC

Tanto en la edad como en el IMC también destacan la monotonía, es decir, al pasar a edades más elevadas la probabilidad de obtener un diagnóstico positivo de diabetes aumenta, al igual que al pasar a mayores niveles de IMC.

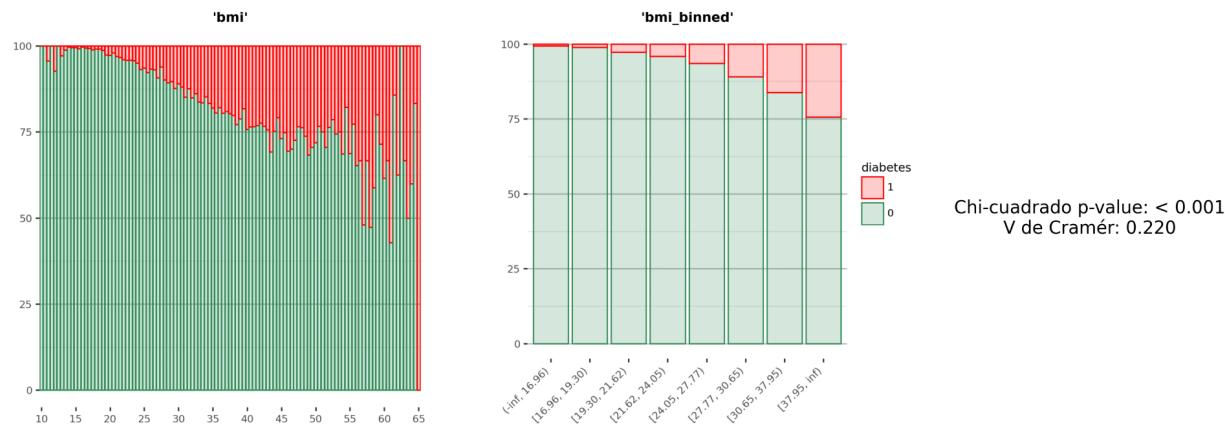
Además de la monotonía observada, los gráficos sugieren que el riesgo de diabetes no aumenta de forma abrupta, sino gradual y progresiva, especialmente en edad, lo que apunta a un efecto acumulativo del envejecimiento. En el caso del IMC, el incremento del riesgo

parece más marcado a partir de los rangos de sobrepeso y obesidad, indicando posibles umbrales de mayor vulnerabilidad clínica.

En conjunto, ambos factores se perfilan como buenos predictores individuales, aunque insuficientes por sí solos para explicar completamente la aparición de diabetes, lo que sugiere que intervienen otros factores adicionales que actúan de forma conjunta.



(a) Distribución de diabetes según la edad



(a) Distribución de diabetes según el IMC

## 3 Test de Hipótesis.

En esta sección describiremos los principales tests estadísticos utilizados para analizar la relación entre variables, tanto categóricas como continuas. Incluiremos algunos de los fundamentos matemáticos y los supuestos de cada prueba y comentaremos brevemente sus utilidades tanto por separado como en conjunto.

### 3.1 Chi-cuadrado ( $\chi^2$ ) y V de Cramer

El test de **Chi-cuadrado** se utiliza para determinar si existe una asociación significativa entre dos variables categóricas. Supongamos que tenemos una tabla de contingencia de  $r$  filas y  $c$  columnas, donde  $O_{ij}$  representa la frecuencia observada en la celda  $(i, j)$  y  $E_{ij}$  la frecuencia esperada bajo independencia:

#### Estadístico $\chi^2$

$$\chi^2 = \sum_{i=1}^r \sum_{j=1}^c \frac{(O_{ij} - E_{ij})^2}{E_{ij}}, \quad E_{ij} = \frac{R_i \cdot C_j}{N}$$

donde  $R_i$  y  $C_j$  son los totales marginales de la fila  $i$  y columna  $j$ , y  $N$  es el tamaño total de la muestra.

#### Hipótesis:

- $H_0$ : Las variables son independientes.
- $H_1$ : Existe dependencia entre las variables.

Para cuantificar la fuerza de la asociación, se puede usar el **V de Cramer**:

#### Estadístico V de Cramer

$$V = \sqrt{\frac{\chi^2}{N \cdot \min(r-1, c-1)}}, \quad 0 \leq V \leq 1$$

Donde valores cercanos a 0 indican poca asociación y valores cercanos a 1 una asociación fuerte. Dependiendo de los valores de las asociaciones que encontramos, se podrá determinar cómo de fuerte es el efecto de una variable categórica en la distribución de otra. Si una de ellas hace más propenso al individuo a desarrollar otras, entonces tendremos una  $V$  grande. Esto además nos permitirá hacer un **ranking** de las variables categóricas que más afectan por ejemplo al diagnóstico de diabetes. Al haber discretizado todas nuestras variables, esto nos permite determinar las asociaciones más fuertes de **todas las variables**.

## 3.2 Test de homogeneidad de varianzas (Levene) y Kruskal-Wallis

El test de Levene se utiliza para comprobar si varios grupos independientes tienen varianzas iguales.

### Estadístico de Levene

Sea  $k$  el número de grupos con tamaños  $n_1, n_2, \dots, n_k$  y  $Y_{ij}$  la observación  $j$  del grupo  $i$ . Se calcula:

$$Z_{ij} = |Y_{ij} - \tilde{Y}_i|$$

donde  $\tilde{Y}_i$  es la mediana (o media) del grupo  $i$ . Luego, se realiza un ANOVA sobre los valores  $Z_{ij}$ :

$$W = \frac{(N - k) \sum_{i=1}^k n_i (\bar{Z}_i - \bar{Z})^2}{(k - 1) \sum_{i=1}^k \sum_{j=1}^{n_i} (Z_{ij} - \bar{Z}_i)^2}$$

Para muestras grandes,  $W$  sigue aproximadamente una distribución  $F$  con  $k - 1$  y  $N - k$  grados de libertad.

### Hipótesis:

- $H_0$ : Las varianzas de todos los grupos son iguales.
- $H_1$ : Al menos un grupo tiene varianza diferente.

### Estadístico de Kruskal-Wallis

Sea  $k$  grupos con tamaños  $n_1, n_2, \dots, n_k$  y un total de  $N = \sum_{i=1}^k n_i$ . Sea  $R_{ij}$  el rango de la observación  $j$  del grupo  $i$  dentro de todos los datos combinados. El estadístico se calcula como:

$$H = \frac{12}{N(N + 1)} \sum_{i=1}^k n_i \left( \bar{R}_i - \frac{N + 1}{2} \right)^2$$

Para  $N$  grande,  $H$  sigue aproximadamente una distribución  $\chi^2$  con  $k - 1$  grados de libertad.

### Hipótesis:

- $H_0$ : Todas las distribuciones de los grupos son iguales.
- $H_1$ : Al menos un grupo difiere de los demás.

El test de **Kruskal-Wallis** compara la distribución de más de dos grupos independientes. En particular, puede utilizarse como test no paramétrico para evaluar la igualdad de medianas entre diversos grupos, sin asumir homogeneidad de varianzas. Además, este test permite comparar múltiples grupos sin asumir normalidad y es especialmente útil cuando los datos presentan valores extremos o distribuciones sesgadas.

### 3.2.1 Kruskal-Wallis y Levene

Aunque Kruskal-Wallis es robusto frente a distribuciones no normales y diferencias de varianza, conocer la homogeneidad de varianzas mediante Levene permite:

- Detectar si las diferencias entre grupos podrían estar acompañadas de heterocedasticidad.
- Ajustar la interpretación de Kruskal-Wallis, ya que grandes diferencias de dispersión pueden influir en la significancia.
- Informar decisiones posteriores sobre modelado o transformaciones de datos, especialmente si se consideran métodos paramétricos.

En la figura 6 se observa que el grupo con diabetes tiene medianas mayores y mayor dispersión. Los p-values de ambos test ( $< 0,001$ ) confirman que existe una diferencia significativa en las distribuciones y que las varianzas no son homogéneas, justificando la evaluación conjunta de ambos aspectos.

## 4

## Insights

### 💡 Insight 1 - Variables a tener en cuenta

Las variables continuas **glucosa en sangre** y **HbA1c** presentan la relación más fuerte con la diabetes, reflejada en un *Information Value (IV)* elevado (Glucosa: 1,58, HbA1c: 1,22). Esto indica que estas variables son altamente discriminativas entre individuos diabéticos y no diabéticos, y serán claves en cualquier modelo predictivo.

### 💡 Insight 2 - Edad e IMC en incidencia de diabetes

La **edad** e **IMC** son factores determinantes, con un IV alto de (1,37) para edad y (0,66) en el caso del IMC. La discretización muestra que individuos menores de 46 años tienen menor riesgo, mientras que aquellos mayores de 54 años presentan mayor probabilidad de diagnóstico. En el caso del IMC, los casos de diabéticos se disparan con un IMC mayor a 27.

### 💡 Insight 3 - Outliers de IMC y riesgo extremo

Un pequeño grupo de individuos con **IMC  $\geq$  65** (47 personas) muestra una prevalencia mucho mayor de diabetes (32 %) y otras comorbilidades. Aunque representan menos del 0,1 % de la población, estos casos extremos podrían influir de manera desproporcionada en modelos predictivos, por lo que se decidió analizarlos por separado y excluirlos del análisis principal.

### 💡 Insight 4 - Relación HbA1c–glucosa en diabetes

En la Figura 7 se observa una relación positiva clara entre **HbA1c** y **glucosa en sangre**, consistente con que ambas métricas capturan el control glucémico. Además, la separación por diagnóstico es marcada: los casos con diabetes se concentran en valores altos de ambas variables, mientras que los no diabéticos se agrupan en rangos más bajos. Visualmente aparecen umbrales prácticos que discriminan bien, con mayor número de diabéticos cuando **blood\_glucose\_level > 200** y **HbA1c\_level > 6,7**. Este patrón sugiere que ambas variables aportan señal complementaria y fuerte para el modelado predictivo.

### 💡 Insight 5 - Perfil clínico según tabaquismo

La incidencia de patologías es similar entre categorías de tabaquismo: **former** muestra la mayor prevalencia y **No Info** la menor. Las diferencias son pequeñas (por lo general  $< 3$  p.p.), salvo en enfermedad cardíaca, donde **ever** aumenta y se aproxima a **former**.

## 5 Metodología de trabajo y organización del equipo

### 5.1 Composición del equipo y organización

El equipo estuvo conformado por cuatro integrantes. Para facilitar el trabajo colaborativo y permitir el desarrollo en paralelo, se estableció una estrategia de control de versiones utilizando GitHub.

La dinámica de trabajo definida fue la siguiente:

- Existencia de una rama principal (`main`) que contenía la versión estable del proyecto.
- Cada integrante desarrollaba sus tareas en una rama independiente (`feature branch`).
- Posteriormente, los cambios eran integrados mediante procesos de `merge` hacia la rama principal.
- Uno de los integrantes asumió la responsabilidad de consolidar los aportes del equipo, resolver conflictos de integración y mantener actualizado un notebook definitivo con la versión final del análisis.

Esta metodología permitió trabajar de forma paralela, ordenada y reproducible.

### 5.2 Código, Entorno de desarrollos, librerías y otras herramientas.

El desarrollo del proyecto se realizó en Python 3.11.9, utilizando notebooks como entorno principal para la ejecución del análisis, la experimentación y la documentación del proceso.

#### 5.2.1 Procesamiento y preparación de datos

- **pandas**: manipulación, limpieza y transformación de datos.
- **NumPy**: operaciones numéricas y manejo eficiente de estructuras matriciales.
- **OptBinning**: discretización y binning óptimo de variables continuas.

#### 5.2.2 Visualización

- **Matplotlib**: generación de gráficos base y visualizaciones personalizadas.
- **plotnine**: construcción de visualizaciones estadísticas basadas en la Gramática de los Gráficos.

### 5.2.3 Análisis estadístico e inferencial

- **SciPy**: pruebas estadísticas y funciones científicas.
- **StatsModels**: modelado estadístico y realización de tests de hipótesis.

### 5.2.4 Estructura del código y reutilización

Con el objetivo de mejorar la modularidad y reutilización del código, se desarrolló un script auxiliar propio denominado `utils.py`. Este archivo centraliza funciones de preprocesamiento, transformaciones frecuentes y rutinas de visualización, tales como *scatter plots*, histogramas y gráficos de barras.

De este modo, el notebook principal mantiene una estructura más limpia y legible, limitándose a importar el módulo (`import utils`) y llamar a las funciones necesarias (`utils.función(argumentos)`), favoreciendo la mantenibilidad, estandarización y reutilización del código.

## 5.3 Flujo de trabajo general

---

El proceso seguido durante el análisis exploratorio se estructuró en las siguientes etapas:

1. Carga e inspección inicial del dataset.
2. Limpieza y tratamiento de valores faltantes o inconsistentes.
3. Transformaciones y discretización de variables relevantes.
4. Análisis descriptivo y visualización de distribuciones mediante gráficos.
5. Evaluación de relaciones entre variables mediante técnicas estadísticas y visuales.
6. Aplicación de pruebas de hipótesis para validar hallazgos.

## 6 Conclusiones y trabajo futuro

Tras realizar el proceso de limpieza, transformación, discretización y análisis exploratorio del dataset, se obtuvo un conjunto de variables estructurado y adecuado para etapas posteriores de modelado predictivo.

El análisis descriptivo y gráfico evidenció que:

- **BMI, edad, HbA1c y glucosa en sangre** presentan una alta capacidad discriminante entre individuos con y sin diabetes, mostrando tendencias monotónicas claras respecto a la variable objetivo.
- Las variables binarias **hipertensión** y **enfermedad cardíaca** se asocian con un aumento significativo en la prevalencia de diabetes, sugiriendo una relación clínica consistente con la literatura.
- El **historial de tabaquismo** y el **género**, aunque menos determinantes de forma individual, aportan información complementaria que podría mejorar el desempeño del modelo al combinarse con otras variables.
- La discretización mediante técnicas de binning permitió capturar relaciones no lineales, reducir la sensibilidad a valores extremos y facilitar la interpretabilidad de los predictores.

En conjunto, estas observaciones confirman que el conjunto de características resultante posee suficiente poder explicativo para abordar el problema como una tarea de **clasificación supervisada binaria**.

### 6.1 Trabajo futuro y modelado predictivo

Como siguiente etapa, se propone la construcción y comparación de múltiples modelos de clasificación, con el objetivo de identificar la alternativa que ofrezca el mejor equilibrio entre desempeño predictivo, robustez e interpretabilidad.

Se plantea evaluar las siguientes familias de modelos:

- **Regresión logística**: modelo base interpretable y adecuado como *baseline*. Permite estimar probabilidades, interpretar coeficientes y analizar la contribución de cada variable mediante *odds ratios*.
- **Árboles de decisión**: capturan relaciones no lineales e interacciones entre variables de forma automática, además de ofrecer reglas fácilmente interpretables.
- **Random Forest**: combinación de múltiples árboles que reduce la varianza y mejora la capacidad de generalización respecto a un árbol individual.
- **Support Vector Machines (SVM)**: útiles para fronteras de decisión no lineales mediante el uso de *kernels*, pudiendo capturar separaciones más complejas del espacio de características.

- **K-Nearest Neighbors (KNN)**: alternativa no paramétrica que puede servir como referencia adicional para comparar desempeño, especialmente tras normalización de variables.

Adicionalmente, se proponen las siguientes buenas prácticas metodológicas:

1. Separación en conjuntos de entrenamiento, validación y prueba.
2. Validación cruzada (*k-fold cross-validation*).
3. Ajuste de hiperparámetros mediante *grid search* o *random search*.
4. Evaluación con múltiples métricas: *accuracy*, *precision*, *recall*, F1-score y AUC-ROC, priorizando *recall* o sensibilidad debido al contexto clínico del problema.

Finalmente, la comparación sistemática de estos enfoques permitirá seleccionar un modelo que no solo maximice el desempeño predictivo, sino que también mantenga un nivel adecuado de interpretabilidad, aspecto especialmente relevante en aplicaciones del ámbito sanitario.

## Referencias Bibliográficas

---

1. World Health Organization. *Diabetes*. Available at: <https://www.who.int/news-room/fact-sheets/detail/diabetes> (Accedido: 2026-01-\*\*).
2. Fundación para la Salud. *Valores objetivos de glucosa y hemoglobina glicosilada*. Disponible en: <https://www.fundacionparalasalud.org/infantil/1049/valores-objetivos-de-glucosa-y-hemoglobina-glicosilada>. (Accedido el: 2026-02-\*\*).
3. World Health Organization. *Nutrition – Maintaining a Healthy Lifestyle*. Disponible en: <https://www.who.int/europe/news-room/fact-sheets/item/nutrition—maintaining-a-healthy-lifestyle>. (Accedido el: 2026-02-\*\*).