UNIVERSIDAD NACIONAL
DE ROSARIO
FACULTAD DE CIENCIAS
ECONÓMICAS Y ESTADÍSTICA

Trabajo Práctico de Modelos Lineales Generalizados

Integrantes:
Candela, Ornella
Mac Kay, Agustina
Ovando, Francisco

Licenciatura en Estadística

Año 2025

Introducción

La diabetes es una enfermedad metabólica crónica caracterizada por niveles elevados de glucosa en sangre, consecuencia de una producción insuficiente de insulina o de una resistencia del organismo a su acción. Esta condición puede derivar en complicaciones cardiovasculares, renales, neurológicas y visuales, afectando de manera significativa la calidad de vida de quienes la padecen.

Dada la relevancia de esta enfermedad en términos de salud pública y su estrecha relación con factores fisiológicos y de estilo de vida, resulta fundamental comprender qué características individuales se asocian a un mayor riesgo de padecerla.

En este trabajo se busca identificar los factores con mayor incidencia en la probabilidad de padecer diabetes, aplicando un modelo lineal generalizado con enlace logit sobre una muestra de mujeres adultas.

Material y métodos

El presente trabajo tiene como objetivo analizar los factores asociados a la presencia de diabetes en 391 mujeres mayores de 21 años, a través de un modelo lineal generalizado con respuesta binaria. De los 391 datos disponibles, la muestra se dividió para el análisis predictivo. Se utilizó un conjunto de entrenamiento de 332 observaciones para ajustar el modelo. Las 59 observaciones restantes se reservaron como conjunto de validación para evaluar la capacidad predictiva. La variable respuesta analizada toma el valor 1 si la persona presenta diagnóstico de diabetes y 0 en caso contrario, por lo que se asume una distribución Bernoulli para cada observación.

Las variables explicativas incluidas en el análisis se describen a continuación:

- **Edad**: medida en años. Es una variable cuantitativa continua.
- **Embarazo**: número de veces que la mujer ha estado embarazada, categorizada en tres niveles:
  - "0": ninguna gestación
  - "1": entre una y tres gestaciones
  - "2": más de tres gestaciones
- **Glucosa**: concentración de glucosa en ayunas. Fue categorizada en dos niveles:
  - "0": valores normales ($\leq$ 100 mg/dL)
  - "1": valores elevados ($>$ 100 mg/dL)
- **Presión arterial**: presión diastólica, clasificada en dos niveles:
  - "0": presión dentro del rango normal ($\leq$ 80 mmHg)
  - "1": presión elevada ($>$ 80 mmHg)

- **Obesidad**: medida a través del índice de masa corporal (IMC). Se definió con dos niveles:
  - "0": IMC menor a 30
  - "1": IMC igual o superior a 30
- **Función de predisposición (DPF)**: índice que cuantifica la predisposición genética a la diabetes a partir del historial familiar. Fue dicotomizado en:
  - "0": valores bajos (≤ 0.5)
  - "1": valores altos (> 0.5)

Análisis descriptivo

A continuación se presenta una visión general de las características clave de la población estudiada.

Tabla 1: Frecuencias absolutas y relativas de diabetes

Diabetes	Frecuencia	Porcentaje (%)
No	261	66.75
Sí	130	33.25
Total	391	100.00

Tabla 2: Frecuencias absolutas y relativas de obesidad

Obesidad	Frecuencia	Porcentaje (%)
Peso Normal o sobrepeso	129	32.99
Obesidad	262	67.01
Total	391	100

Tabla 3: Frecuencias absolutas y relativas de glucosa

Glucosa	Frecuencia	Porcentaje (%)
Normal	116	29.67
Elevada	275	70.33
Total	391	100

Tabla 4: Frecuencias absolutas y relativas de presión

Presión	Frecuencia	Porcentaje (%)
Normal	313	80.05
Elevada	78	19.95
Total	391	100

Tabla 5: Frecuencias absolutas y relativas de DPF

DPF	Frecuencia	Porcentaje (%)
Riesgo bajo	223	57.03
Riesgo alto	168	42.97
Total	391	100

Tabla 6: Frecuencias absolutas y relativas de embarazo

Embarazos	Frecuencia	Porcentaje (%)
Ninguno	56	14.32
1 a 3	202	51.66
Más de 3	133	34.02
Total	391	100

Tabla 7: Estadísticas descriptivas de la variable edad

Variable	Media	Mediana	DE	Mínimo	Máximo
Edad	30.74	27	9.89	21	63

Los datos muestran que un 33.2% de la población estudiada tiene diabetes, con una alta prevalencia de obesidad (67%) y niveles elevados de glucosa (70.3%). Además, el 43% de los participantes tiene antecedentes familiares de diabetes, lo que podría indicar una predisposición genética relevante. Aunque la mayoría presenta presión arterial normal (80.1%), los factores como la obesidad, los niveles de glucosa y el historial familiar podrían ser más críticos para entender el riesgo de diabetes en esta población.

Seguidamente, se presentan gráficos que permiten observar la relación entre la variable respuesta (presencia de diabetes) y distintas variables explicativas categóricas y continuas. Esta visualización facilita la detección de posibles asociaciones y patrones relevantes para el posterior análisis.

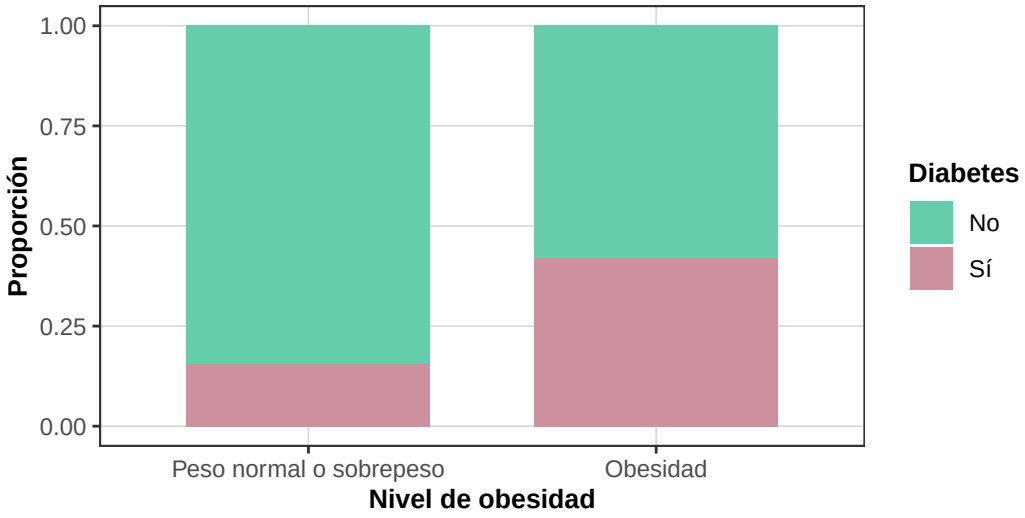

Figura 1: Proporción de personas con diabetes según nivel de obesidad

Se observa que la proporción de mujeres con diabetes es más alta para aquellas con obesidad en comparación a las pacientes sin obesidad.

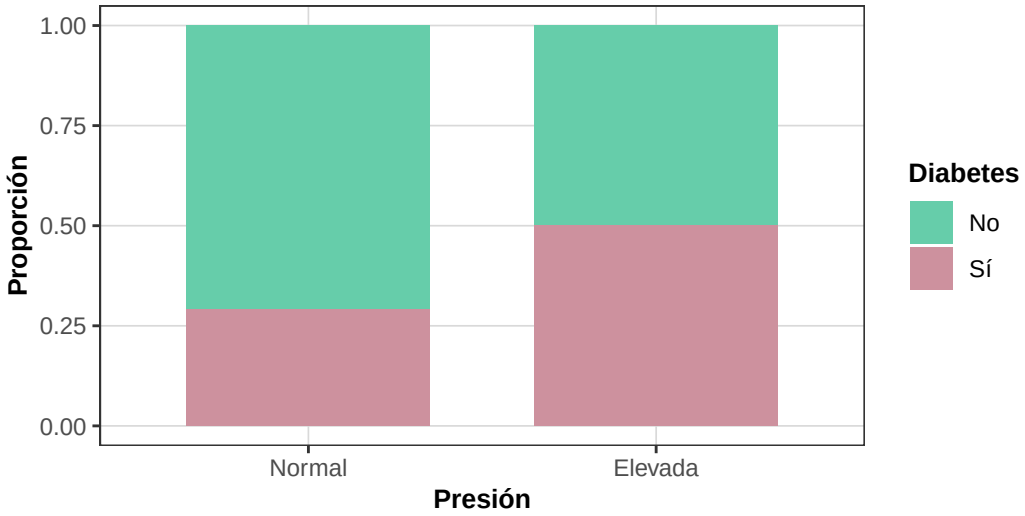

Figura 2: Proporción de personas con diabetes según presión

La presión elevada puede estar influyendo en la probabilidad de padecer diabetes, dado que se puede ver una proporción más alta de diabéticas en las mujeres con presión elevada que en aquellas con presión normal.

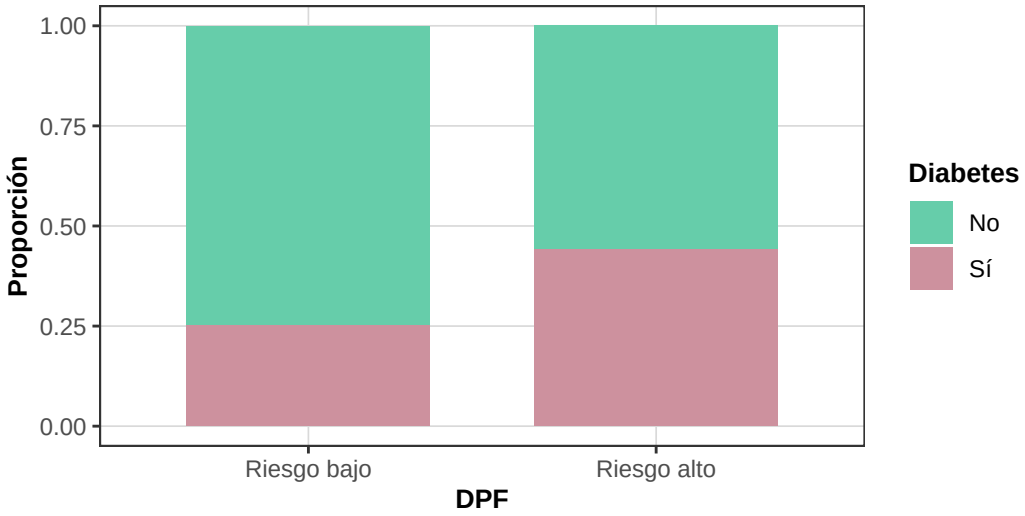

Figura 3: Proporción de personas con diabetes según DPF

Como es de esperarse, hay mayor proporción de diabéticas en aquellas mujeres con un DPF más alto comparando con las pacientes con DPF bajo.

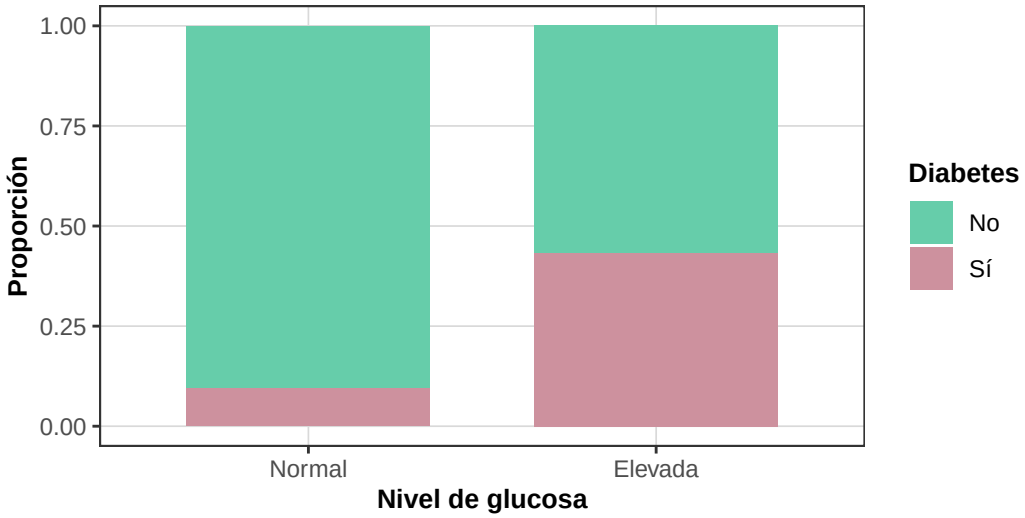

Figura 4: Proporción de personas con diabetes según nivel de glucosa

La relación entre la glucosa y la diabetes es directa por la naturaleza de la enfermedad, hay mayor proporción de diabéticas en pacientes con glucosa elevada que en aquellas con niveles de glucosa normales.

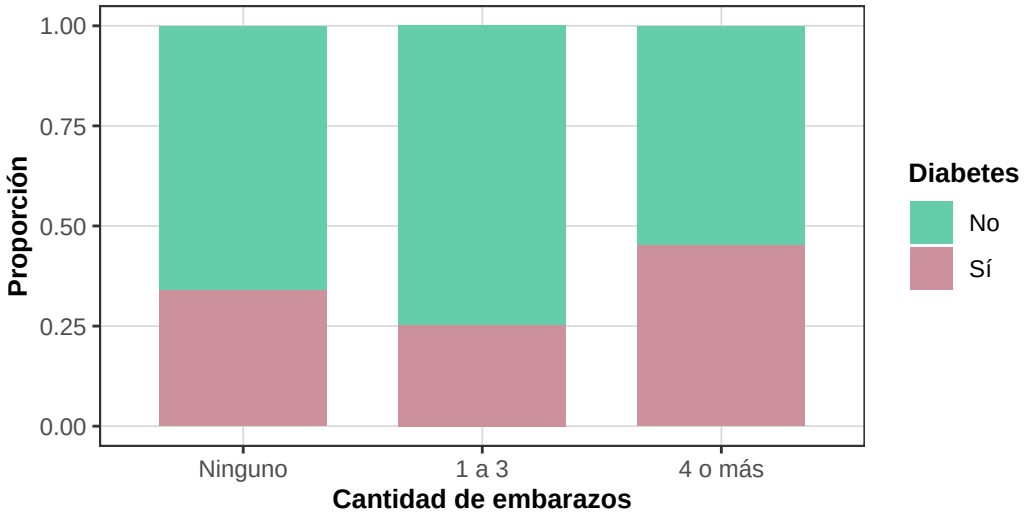

Figura 5: Proporción de personas con diabetes según cantidad de embarazos

Parece haber un aumento en la proporción de mujeres con diabetes en aquellas que tuvieron 4 o más embarazos, y un descenso para las que tuvieron entre 1 y 3, en comparación a las mujeres que no tuvieron ningún embarazo.

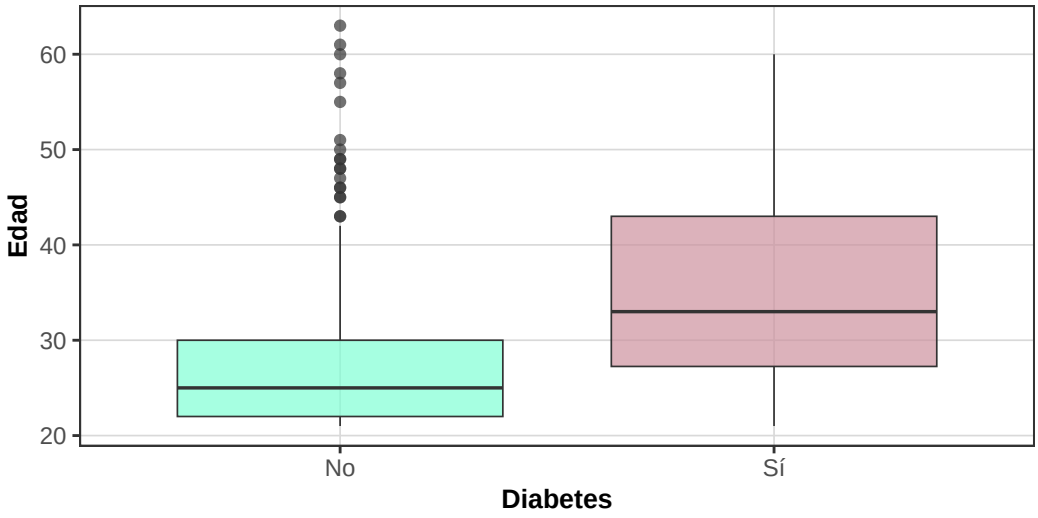

Figura 6: Distribución de la edad según presencia de diabetes

El 50% de las mujeres sin diabetes parece tener entre 23 y 30 años aproximadamente, mientras que el 50% de las mujeres con diabetes tienen entre 27 y 43 años aproximadamente. Esto habla de que las mujeres con diabetes en general tienen una edad mayor a las mujeres sin diabetes.

Modelado estadístico

El trabajo se realizará bajo el enfoque de modelos lineales generalizados, utilizando lo desarrollado por Nelder y Wedderburn.

Selección del modelo

Para modelar la probabilidad de presentar diabetes en mujeres mayores a 21 años, en función de las variables explicativas edad, glucosa, presión arterial, DPF, embarazo y obesidad, se propuso emplear un modelo con función de enlace logit.

En primer lugar, se aplicaron los métodos de selección de variables hacia adelante y hacia atrás, eligiendo en cada paso el modelo con menor medida de la información de Akaike (AIC). En ambos casos se obtuvo el mismo modelo final, compuesto únicamente por 4 efectos principales: los de edad, glucosa, DPF y obesidad.

A continuación, se exploró la posibilidad de incorporar interacciones de segundo orden entre las variables explicativas, ajustando un modelo ampliado y comparándolo con el modelo sin interacciones. Dado que la inclusión de interacciones no mejoró el ajuste, se mantuvo el modelo de efectos principales.

Con el objetivo de verificar el supuesto de linealidad entre la variable continua edad y el logit de la probabilidad de presentar diabetes, se implementó una comparación entre dos modelos alternativos.

En primer lugar, se construyó un modelo que incorpora la variable edad como ordinal, dividiendo su rango en cinco categorías según los cuantiles 0.2, 0.4, 0.6 y 0.8. Posteriormente, se ajustó un segundo modelo en el que la edad se representó mediante variables indicadoras (dummies), permitiendo una relación no necesariamente lineal con la respuesta.

El modelo ordinal se puede expresar como:

$$\text{logit}(\pi_i) = \beta_0 + \beta_1 \,\text{Glucosa}_i + \beta_2 \,\text{Obesidad}_i + \beta_3 \,\text{DPF}_i + \beta_4 \,\text{EdadORDEN}_i$$

Mientras que el modelo con variables indicadoras se define como:

$$\text{logit}(\pi_i) = \beta_0 + \beta_1 \,\text{Glucosa}_i + \beta_2 \,\text{Obesidad}_i + \beta_3 \,\text{DPF}_i + \beta_4 \,\text{Edad2}_i + \beta_5 \,\text{Edad3}_i + \beta_6 \,\text{Edad4}_i + \beta_7 \,\text{Edad5}_i$$

Ambos modelos fueron comparados mediante la prueba de razón de verosimilitud. Dado que la diferencia en el ajuste entre el modelo con edad tratada como ordinal y el modelo con las variables dummies no resultó estadísticamente significativa, se concluye que la relación entre la edad y el logit de la probabilidad de diabetes puede considerarse lineal ($p-value = 0.6408$). Por ende, se mantiene el efecto lineal de la variable edad en el modelo.

Por lo tanto, el modelo seleccionado se expresa como:

$$\text{logit}(\pi_i) = \ln\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 \,\text{Edad}_i + \beta_2 \,\text{Glucosa}_i + \beta_3 \,\text{DPF}_i + \beta_4 \,\text{Obesidad}_i, \quad i = \overline{1,332}$$

donde:

- $\text{Glucosa}_i = \begin{cases} 1 & \text{si Glucosa} > 100 \\ 0 & \text{si Glucosa} \leq 100 \end{cases}$
- $\text{DPF}_i = \begin{cases} 1 & \text{si Diabetes Pedigree Function} > 0.5 \\ 0 & \text{si Diabetes Pedigree Function} \leq 0.5 \end{cases}$
- $\text{Obesidad}_i = \begin{cases} 1 & \text{si IMC} \geq 30 \\ 0 & \text{si IMC} < 30 \end{cases}$

Bondad del ajuste y estimación del modelo

Se evalúa la adecuación de la función de enlace logit utilizada en el modelo aplicando el test de especificación de la función de enlace, que consiste en incorporar al modelo un término adicional cuadrático del predictor lineal estimado:

$$M1) \quad \text{logit}(\pi_i) = \beta_0 + \beta_1 \,\text{Edad}_i + \beta_2 \,\text{Glucosa}_i + \beta_3 \,\text{DPF}_i + \beta_4 \,\text{Obesidad}_i \quad i = \overline{1,332}$$
$$M2) \quad \text{logit}(\pi_i) = \beta_0 + \beta_1 \,\text{Edad}_i + \beta_2 \,\text{Glucosa}_i + \beta_3 \,\text{DPF}_i + \beta_4 \,\text{Obesidad}_i + \beta_5 \,\hat{\eta}_i^2 \quad i = \overline{1,332}$$

El procedimiento implica ajustar un modelo extendido que incluye esta nueva variable y comparar su ajuste con el modelo original mediante una prueba de razón de verosimilitud. En este análisis, la comparación entre ambos modelos no mostró evidencia suficiente para rechazar la adecuación del enlace logit, por lo que se considera apropiado mantenerlo en el modelo final ($p - value = 0.6489$).

Además, para evaluar la bondad de ajuste del modelo, se aplicó la prueba de Hosmer y Lemeshow. Esta prueba compara las frecuencias observadas y esperadas del evento de interés en grupos formados a partir de los valores predichos por el modelo y dado el tamaño muestral disponible, se optó por dividir las observaciones en 10 grupos de igual tamaño según los valores de probabilidad predicha. Finalmente, se concluye que el modelo presenta un ajuste adecuado a los datos ($p - value = 0.7069$).

Considerando los resultados del estudio del modelo a través de los diferentes tests de hipótesis, el ajuste realizado es el siguiente:

$$\text{logit}(\hat{\pi}_i) = \ln\left(\frac{\hat{\pi}_i}{1-\hat{\pi}_i}\right) = -6.04 + 0.07 \cdot \text{Edad}_i + 1.91 \cdot \text{Glucosa}_i + 0.98 \cdot \text{DPF}_i + 1.47 \cdot \text{Obesidad}_i \quad i = \overline{1,332}$$

Análisis de residuos

El análisis de residuos permite evaluar la adecuación del modelo ajustado, tanto en lo que respecta a la componente sistemática como a la aleatoria.

Con el objetivo de analizar la estructura del modelo, se construyó el gráfico de los residuos cuantil frente a las probabilidades estimadas bajo el enlace logit.

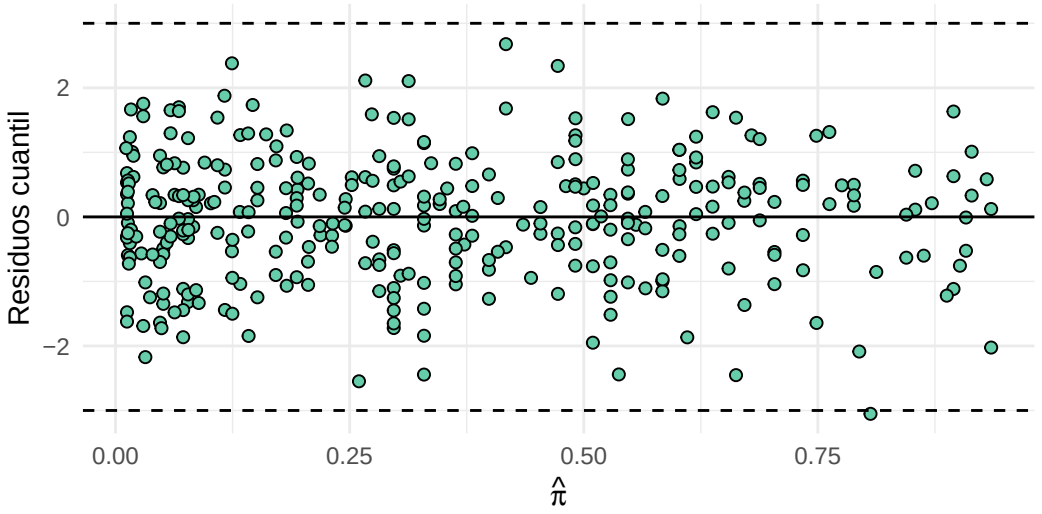

Figura 7: Gráfico de residuos cuantil vs. las probabilidades estimadas

En este gráfico, la mayoría de los puntos se distribuyen de manera aleatoria alrededor de la línea horizontal en cero, sin evidenciar un patrón sistemático.
Esto sugiere que la función de enlace logit es adecuada y que el modelo logra capturar correctamente la relación entre las variables explicativas y la probabilidad de presentar diabetes.

Además, para verificar la adecuación de la distribución supuesta de los errores, se empleó un gráfico QQ de los residuos cuantil.

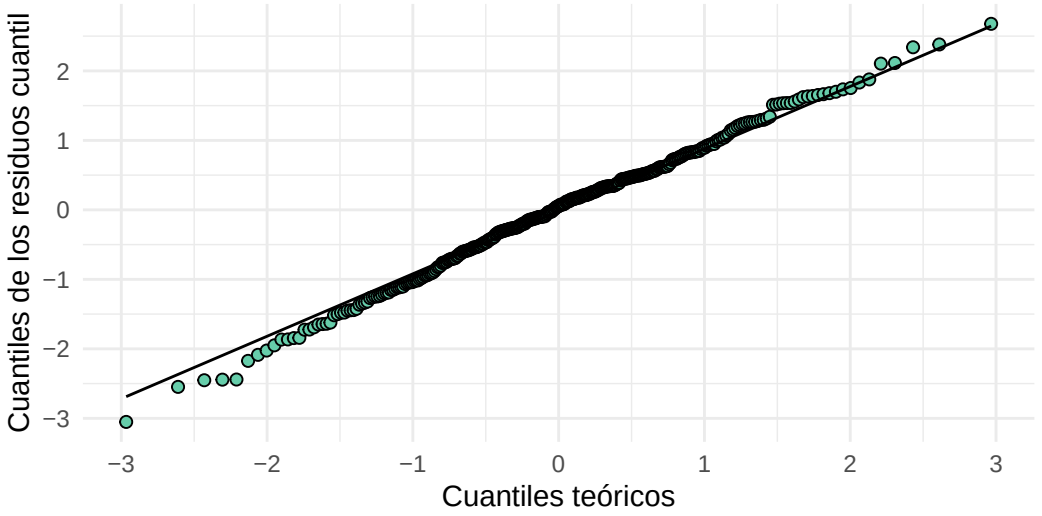

Figura 8: Gráfico probabilístico normal con residuos cuantil

En dicho gráfico se observa que los puntos siguen aproximadamente la línea de referencia, con ligeras desviaciones en los extremos.
Por lo tanto, puede concluirse que no existen desviaciones importantes respecto a la

distribución esperada, indicando un ajuste razonablemente bueno del modelo.

Estudio de la capacidad predictiva del modelo

Con el objetivo de evaluar la capacidad predictiva del modelo ajustado, se procedió a dividir la base de datos en dos subconjuntos: un conjunto de entrenamiento, utilizado para estimar los parámetros del modelo, y un conjunto de validación, empleado para evaluar su desempeño sobre observaciones no utilizadas en el ajuste. Esta separación permite obtener una medida más realista de la capacidad del modelo para generalizar a nuevos datos.

A partir de las probabilidades estimadas en el conjunto de validación, se construyó la curva ROC, que representa la sensibilidad y especificidad frente a para distintos puntos de corte. Este gráfico permite analizar el equilibrio entre verdaderos positivos y falsos positivos, siendo un instrumento útil para determinar el umbral óptimo de clasificación.

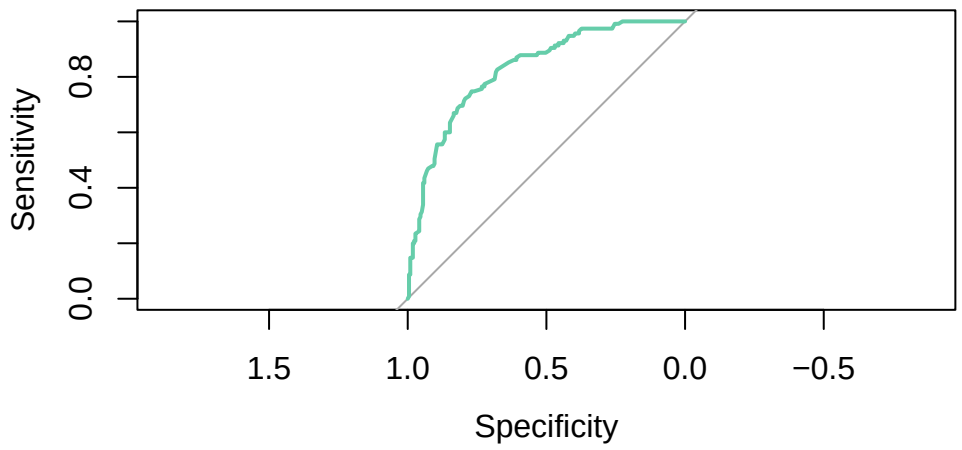

Figura 9: Curva ROC

El punto de corte seleccionado fue de 0.377, correspondiente al valor que maximiza simultáneamente la sensibilidad y la especificidad del modelo. Con este umbral se obtuvo la siguiente matriz de confusión, que resume el desempeño del modelo en la clasificación de los casos:

Predicho / Observado	**No diabetes**	**Diabetes**
No diabetes	31	7
Diabetes	10	11

El modelo logró clasificar correctamente al 71% de las observaciones del conjunto de validación, con una sensibilidad del 61% (proporción de diabéticas correctamente identificadas) y una especificidad del 75.6% (proporción de no diabéticas correctamente clasificadas).

Resultados

En esta sección se presentan los principales hallazgos obtenidos a partir del modelo final ajustado. En primer lugar, se exponen las razones de odds estimadas para cada variable significativa, seguidas por la representación gráfica de las mismas. Posteriormente, se incluyen las curvas ajustadas que ilustran la probabilidad predicha de diabetes bajo diferentes combinaciones de factores.

Razones de odds

Las razones de odds permiten cuantificar la variación en la probabilidad de padecer diabetes ante cambios en las variables explicativas, manteniendo constantes las demás condiciones.

Tabla 9: Razones de Odds estimadas por el modelo

		IC RO	
Variable	$\widehat{RO}$	Límite inferior	Límite superior
Glucosa (elevada vs. normal)	6.77	3.24	15.68
Obesidad (sí vs. no)	4.33	2.25	8.85
DPF (alto vs. bajo)	2.65	1.54	4.62
Edad (cambio en 5 años)	1.46	1.26	1.69

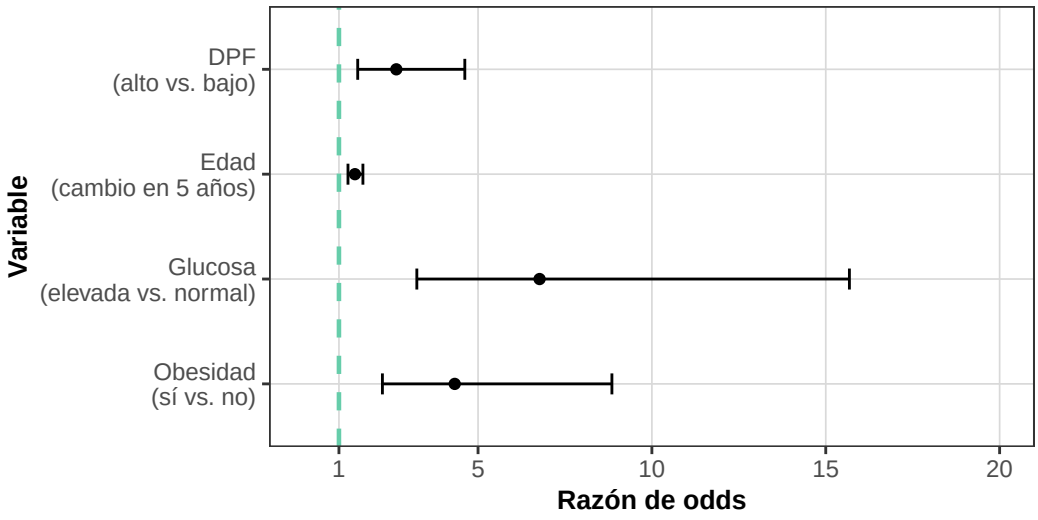

Figura 10: Forest plot de Razones de odds (cambios en edad por 5 años)

De la Tabla 9 y la Figura 10 se extrae que:

- La chance de tener diabetes para las mujeres con glucosa elevada es entre 3.24 y 15.68 veces la misma chance para mujeres con niveles de glucosa normales, para niveles fijos de obesidad, DPF y misma edad.
- La chance de tener diabetes para mujeres con obesidad es entre 1.25 y 7.85 veces mayor a la misma chance para mujeres sin obesidad, con niveles fijos de glucosa, dpf y misma edad.
- La chance de tener diabetes para mujeres con dpf alto es entre un 54% y un 362% mayor que la misma chance para mujeres con dpf bajo, para niveles fijes de glucosa, obesidad y misma edad.
- La chance de tener diabetes aumenta entre un 26% y un 69% al aumentar la edad de la mujer en 5 años, para niveles fijos de glucosa, obesidad y dpf.

Probabilidades ajustadas

Las curvas de probabilidades ajustadas permiten visualizar cómo se modifica el riesgo estimado para los diferentes perfiles en estudio y la tendencia creciente de la probabilidad de diabetes con la edad.

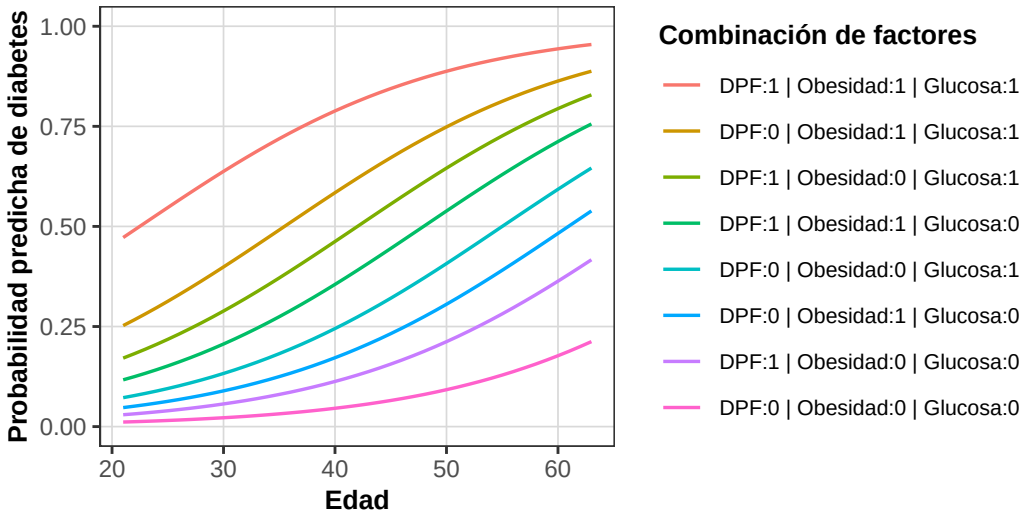

Figura 11: Probabilidad predicha de diabetes según edad y combinación de factores

Dado que el modelo no contempla interacciones, se observa que para todos los perfiles hay un aumento en la probabilidad de tener diabetes al aumentar la edad de la paciente. Además, se destaca que la curva que está por debajo de las restantes corresponde a las mujeres sin obesidad, con niveles de glucosa normales y un DPF normal, mientras que la curva por encima de las otras, es la de las mujeres con obesidad, niveles de glucosa elevados y un DPF alto.

Conclusión

El presente trabajo permitió ajustar y evaluar un modelo lineal generalizado con enlace logit para estudiar los factores asociados a la presencia de diabetes en una muestra de mujeres adultas. A través de un proceso de selección, diagnóstico y validación, se obtuvo un modelo parsimonioso y con un buen nivel de ajuste según las pruebas aplicadas.

El análisis evidenció que los niveles de glucosa en ayunas, la obesidad, los antecedentes familiares (DPF) y la edad contribuyen significativamente a explicar la probabilidad de padecer diabetes. La validación del modelo, mediante la curva ROC y la matriz de confusión, mostró una capacidad predictiva adecuada, confirmando su utilidad para clasificar correctamente una proporción importante de los casos.