REPORTE TÉCNICO / MACHINE LEARNING

Estimación del riesgo de diabetes en pacientes mexicanos

Un modelo para priorizar pruebas confirmatorias de diabetes sin usar glucosa ni HbA1c, con validación por hogar, análisis de sesgos y plan de operación.

Alan Adelfo Del Angel Figueroa
Ingeniería en Desarrollo de Software, Universidad Tecmilenio
Proyecto final de Algoritmos de Aprendizaje Automático / Septiembre de 2026
Código: github.com/AlanDelAngel/diabetes_pacientes_mexicanos

Resumen

Este proyecto construye un modelo de aprendizaje automático que ordena a pacientes adultos mexicanos según su probabilidad de tener riesgo alto de diabetes, usando datos que suelen estar disponibles antes de una prueba de glucosa: edad, sexo, peso, estatura y un perfil bioquímico (lípidos, insulina, ácido úrico, creatinina y albúmina). El objetivo es decidir a quién conviene hacerle primero una prueba confirmatoria (glucosa en ayuno o HbA1c) cuando ese recurso es limitado. **El modelo no diagnostica**: apoya una decisión que siempre revisa personal de salud.

El modelo final es una máquina de soporte vectorial (SVM) con kernel RBF, ajustada por búsqueda de hiperparámetros y evaluada con validación cruzada agrupada por hogar. En un conjunto de prueba de 365 pacientes de hogares que el modelo nunca vio, con un umbral de 0.26 elegido para garantizar un recall de al menos 80%, obtiene:

0.825	**0.765**	**0.481**	**0.764**
Recall de riesgo alto (99 de 120 casos detectados)	ROC-AUC en prueba (0.505 el modelo de referencia)	Precisión: aceptable en un esquema de tamizaje con revisión	ROC-AUC en validación cruzada (+/- 0.027, 15 pliegues)

El desempeño es parejo entre hombres y mujeres (recall de 0.83 y 0.82), pero baja en adultos de 20 a 39 años (0.47 en prueba). Esa limitación queda documentada en la ficha del modelo y en el registro de riesgos.

> **Cómo usar este reporte con el repositorio.** Todo el análisis está en el notebook `Proyecto_Final_Riesgo_Diabetes.ipynb`. Cada sección de este reporte indica la sección del notebook donde se generan sus cifras y figuras. Al ejecutarlo, el notebook crea las carpetas `scripts/`, `artifacts/`, `results/` y `docs/` con los archivos que se mencionan aquí (sección 11).

Contenido

1. Problema y criterio de éxito
2. Datos y calidad
3. Diseño de la validación
4. Pipeline e ingeniería de características
5. Comparación y ajuste de modelos
6. Umbral de decisión
7. Resultados en el conjunto de prueba
8. Interpretabilidad
9. Errores, sesgos y uso responsable
10. Puesta en operación
11. Reproducibilidad
12. Ficha del modelo

1. Problema y criterio de éxito

La variable objetivo es `riesgo_diabetes_cat`, con tres niveles: 0 = bajo, 1 = moderado y 2 = alto. La decisión operativa es binaria: identificar a los pacientes de **riesgo alto** frente al resto, porque son quienes más se benefician de una prueba confirmatoria temprana.

Criterio de éxito: detectar el riesgo alto sin usar glucosa ni HbA1c, con un ROC-AUC claramente superior al de un modelo de referencia (`DummyClassifier`) y un recall de riesgo alto de al menos 0.80. Se prioriza el recall porque, en salud, no detectar a un paciente de riesgo alto cuesta más que una prueba confirmatoria adicional.

2. Datos y calidad

Notebook: Parte I (carga, glosario, análisis exploratorio y calidad de datos).

Se usó el *Diabetes Mexico Data Set* (Félix Jiménez y Sánchez Lee, 2026), derivado de una encuesta nacional de salud tipo ENSANUT: **2,549 adultos de 2,055 hogares** y 24 columnas. El objetivo está muy desbalanceado: 1,697 pacientes de riesgo bajo (66.6%), 12 de riesgo moderado (0.5%) y 840 de riesgo alto (33.0%).

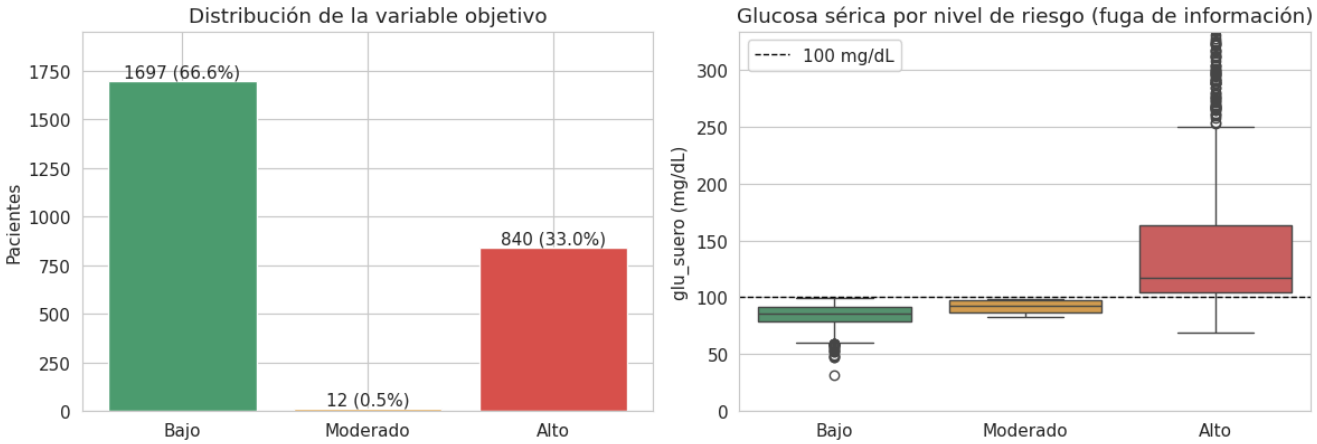


Figura 1. Distribución de la variable objetivo (izquierda) y glucosa sérica por nivel de riesgo (derecha), que muestra por qué la glucosa produce fuga de información.

Fuga de información

La glucosa sérica (`glu_suero`) y la HbA1c (`hb1ac`) son las variables con las que se construyó el objetivo. Incluirlas daría un modelo casi perfecto en papel e inútil en la práctica, porque el propósito es justamente priorizar a quién medirle la glucosa. Por eso **se excluyeron**, igual que los índices TyG y HOMA-IR, que también dependen de la glucosa.

Calidad de los datos

- **Valores centinela:** 11 registros con peso = 222.22 y 15 con estatura = 222.2 generaban IMC ficticios cercanos a 45. Se trataron como faltantes.
- **Faltantes:** 28% de los pacientes no tiene peso o estatura; la columna `imc` original tiene 98.6% de faltantes y se reemplazó por un IMC calculado.
- **Sin duplicados**; `folio_int` identifica de forma única a cada paciente.

Variables excluidas	Motivo
`glu_suero`, `hb1ac`	Fuga de información (definen el objetivo)
`folio_i`, `folio_int`	Identificadores (el folio de hogar se usa solo para agrupar la validación)
`Ciudad`, `entidad_cve`	Alta cardinalidad; el modelo sin entidad rinde igual y es más interpretable (sección 5)
`imc`, `muestra_suero`	98.6% de faltantes; variable constante
`ponde_venosa`, `estrato`, `est_sel`, `ponde_hemo`	Variables del diseño muestral de la encuesta

3. Diseño de la validación

Notebook: Parte I (partición por hogar) y Parte II, sección 6.

- **Partición por hogar:** los integrantes de un mismo hogar comparten alimentación, actividad física y entorno. Con `StratifiedGroupKFold` (semilla 42) se separó un conjunto de **desarrollo de 2,184 pacientes** y uno de **prueba de 365**, sin ningún hogar compartido.
- **Validación cruzada repetida:** 5 pliegues agrupados por hogar, repetidos con 3 semillas (42, 43 y 44), para un total de 15 evaluaciones por modelo. Una sola partición era inestable: con tan pocos casos de riesgo moderado, un pliegue podía tener uno o ninguno.
- **Métrica principal:** ROC-AUC de riesgo alto frente al resto. Secundarias: recall, precisión y F1 de riesgo alto, accuracy y F1 macro.
- **Conjunto de prueba intacto:** no se usó para comparar modelos, elegir variables, ajustar hiperparámetros ni elegir el umbral. Solo se usó una vez, en la sección 7.

4. Pipeline e ingeniería de características

Notebook: Parte II, secciones 4 y 5. Código del transformador: `scripts/diabetes_transformers.py`.

Todo el preprocesamiento vive en un único `Pipeline` de scikit-learn, para que entrenamiento e inferencia apliquen exactamente las mismas reglas:

1. `ConstructorCaracteristicas`: limpieza de centinelas y creación de variables nuevas.
2. `ColumnTransformer` con cuatro ramas: numéricas (winsorización p0.5-p99.5, imputación por mediana y escalamiento), numéricas sesgadas (lo mismo más `log1p`), binarias (One-Hot) y ordinales (`OrdinalEncoder`).
3. El modelo.

Todos los parámetros (límites de winsorización, medianas, categorías) se ajustan solo con datos de entrenamiento de cada pliegue.

Variable nueva	Construcción	Interpretación clínica	Riesgo de fuga
`ratio_trig_hdl`	trig / col_hdl	Marcador indirecto de resistencia a la insulina	Bajo
`col_no_hdl`	colest - col_hdl	Colesterol aterogénico (LDL + VLDL)	Bajo
`grupo_edad`	20-39, 40-59, 60+	Etapas de vida con riesgo creciente	Nulo
`categoria_imc`	<25, 25-29.9, >=30 (OMS)	Peso normal, sobrepeso, obesidad	Nulo

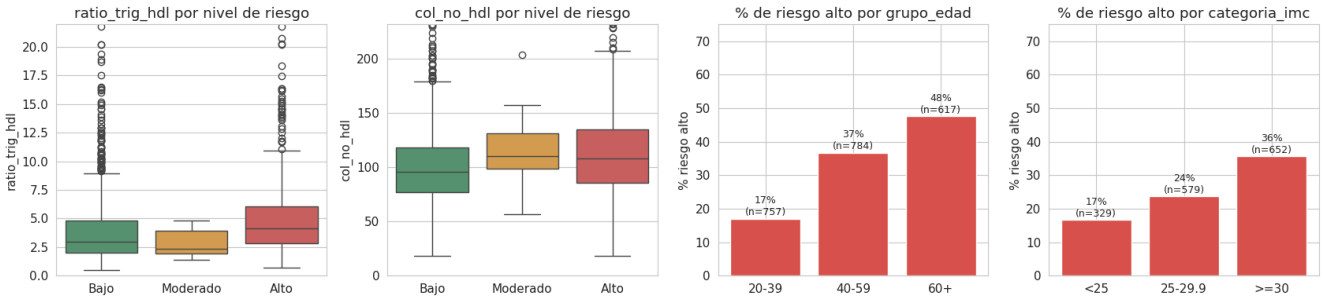


Figura 2. Relación de las variables nuevas con el nivel de riesgo (datos de desarrollo).

5. Comparación y ajuste de modelos

Notebook: Parte II, secciones 7 a 14. Registro completo: `results/registro_experimentos.csv`.

5.1 Familias de modelos

Se compararon ocho modelos con el mismo pipeline y la misma validación cruzada (valores con la configuración por defecto de cada modelo):

Modelo	Familia	ROC-AUC alto	Desv.	Recall alto	F1 alto	Inferencia (ms/1000)
DummyClassifier	Referencia	0.505	0.026	0.338	0.337	22.2
Regresión logística	Lineal	0.723	0.022	0.642	0.566	17.5
Árbol de decisión	Árbol	0.674	0.033	0.608	0.525	27.7
Random Forest	Ensamble	0.756	0.026	0.579	0.561	72.3
Gradient Boosting (HGB)	Ensamble	0.746	0.026	0.642	0.571	35.0
SVM lineal	Márgenes	0.752	0.023	0.684	0.576	47.8
SVM RBF	**Márgenes**	**0.764**	**0.026**	**0.734**	**0.601**	**133.0**
Red neuronal (MLP)	Red neuronal	0.720	0.033	0.306	0.378	16.6

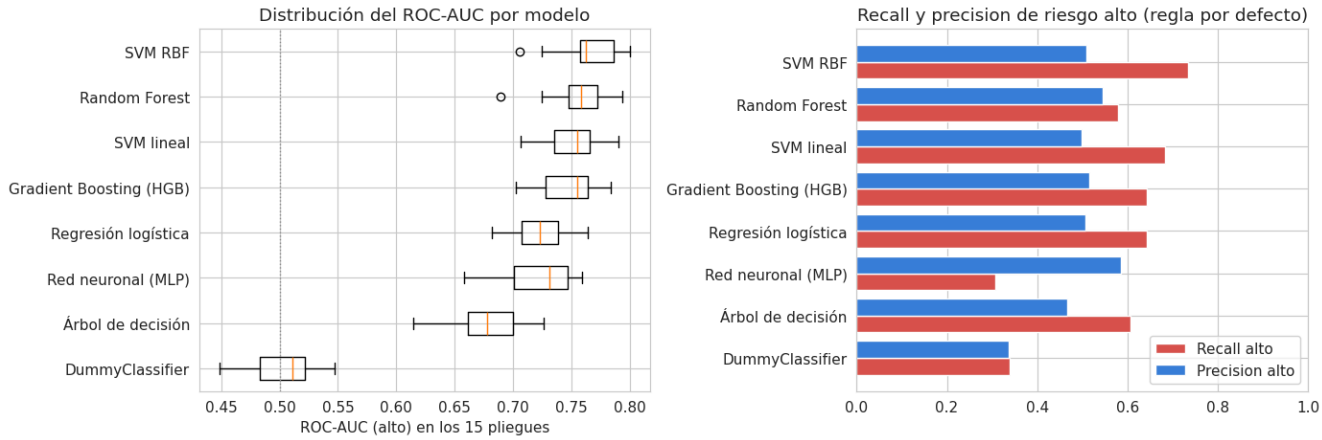


Figura 3. ROC-AUC de riesgo alto en los 15 pliegues (izquierda) y recall y precisión con la regla por defecto (derecha).

5.2 Selección de características

Se probaron dos enfoques dentro de la validación cruzada: `SelectFromModel` con importancias de Random Forest y un modelo reducido sin la entidad federativa (15 variables clínicas y demográficas).

Configuración	Características	ROC-AUC alto	Desv.	Interpretabilidad
SVM RBF, modelo completo	41	0.764	0.026	Media: incluye entidades One-Hot
SVM RBF + SelectFromModel	21	0.761	0.027	Media: la selección cambia en cada reentrenamiento
SVM RBF, modelo reducido	**15**	**0.762**	**0.028**	**Alta: solo variables clínicas y demográficas**
Random Forest, modelo completo	41	0.756	0.026	Media
Random Forest, modelo reducido	15	0.755	0.027	Alta

El modelo reducido prácticamente iguala al completo (diferencia de 0.002 o menos en ROC-AUC), elimina las columnas regionales de la entidad federativa, difíciles de explicar clínicamente, y es más barato de operar. Se adopta en adelante.

5.3 Ajuste de hiperparámetros

Modelo	Espacio explorado	Combinaciones	Mejor configuración	ROC-AUC alto
SVM RBF (GridSearchCV)	**C {0.1, 0.3, 1, 3, 10} x gamma {scale, 0.01, 0.03, 0.1}**	**20**	**C = 3, gamma = 0.03**	**0.764**
SVM lineal (GridSearchCV)	C {0.01, 0.03, 0.1, 0.3, 1}	5	C = 0.1	0.747
Random Forest (RandomizedSearchCV)	n_estimators, max_depth, min_samples_leaf, max_features	20	400 árboles, min_samples_leaf = 3, max_features = 0.3	0.757

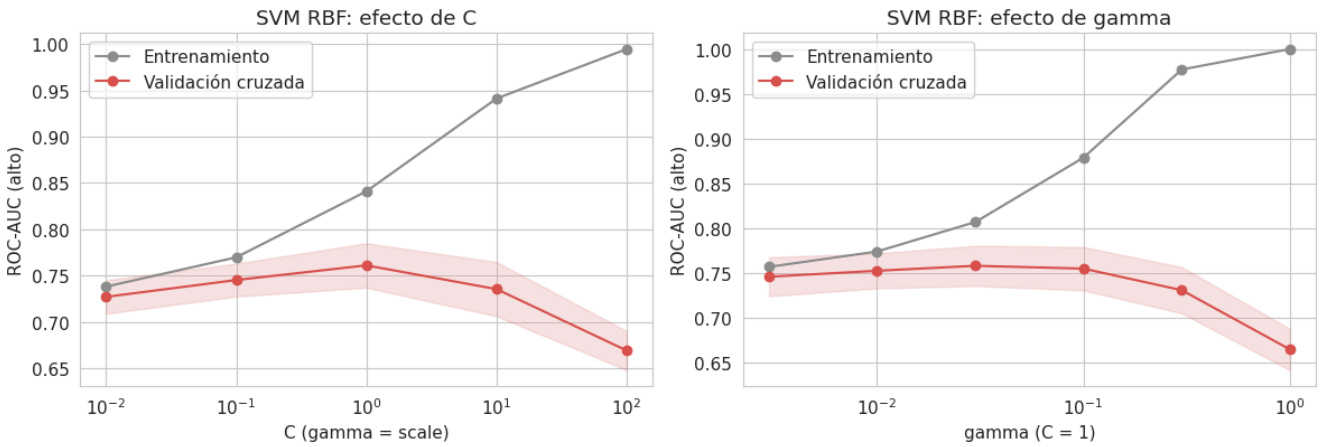


Figura 4. Curvas de validación de la SVM RBF: con valores más altos de C o gamma sube el ROC-AUC de entrenamiento pero baja el de validación (sobreajuste).

5.4 Matriz de decisión entre finalistas

Se eligieron como finalistas los dos mejores modelos ajustados de familias distintas, ambos con el pipeline reducido. La decisión no se tomó solo por la métrica más alta: cada criterio se puntuó de 1 a 5 con evidencia de la validación cruzada.

Criterio	Peso	SVM RBF	Random Forest	Evidencia (validación cruzada)
Recall	25%	5.0	2.64	0.710 frente a 0.491 (referencia 0.338)
ROC-AUC	20%	5.0	4.91	0.764 frente a 0.758 (referencia 0.505)
F1-score	15%	5.0	3.98	0.590 frente a 0.526
Precisión	10%	3.92	5.0	0.506 frente a 0.569
Estabilidad	10%	4.9	5.0	Desviación del ROC-AUC: 0.027 frente a 0.026
Interpretabilidad	10%	3.0	4.0	RF admite SHAP exacto; la SVM solo métodos agnósticos
Costo operativo	10%	4.0	5.0	Inferencia 82 frente a 65 ms/1000; artefacto 0.27 frente a 18.16 MB
Total ponderado	**100%**	**4.58**	**4.14**	**La ventaja en recall, el criterio de mayor peso, decide**

6. Umbral de decisión

Notebook: Parte II, sección 15. Archivo: `artifacts/umbral_v1.json`.

El umbral se eligió con predicciones fuera de pliegue del conjunto de desarrollo (cada paciente evaluado por un modelo que no lo vio en su entrenamiento), sin tocar el conjunto de prueba. Regla: entre los umbrales que detectan al menos 80% de los casos de riesgo alto, se elige el más alto, que produce menos falsos positivos. Como referencia se calcularon el índice de Youden y el umbral de costo mínimo (un falso negativo cuesta 3 veces lo que un falso positivo).

Umbral	Precisión	Recall	F1	Pacientes señalados	Criterio
0.21	0.448	0.882	0.595	64.8%	Costo mínimo
0.26	**0.470**	**0.813**	**0.595**	**57.0%**	**Elegido: recall >= 0.80**
0.30	0.495	0.756	0.598	50.3%	
0.37	0.539	0.661	0.594	40.4%	Índice de Youden
0.50	0.587	0.440	0.503	24.7%	Regla por defecto

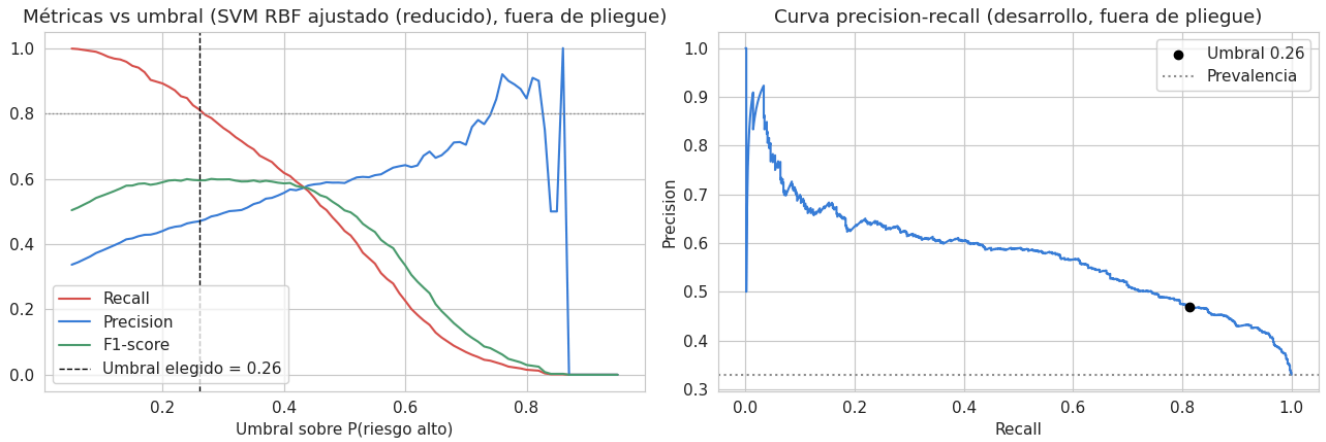


Figura 5. Efecto del umbral sobre precisión, recall y F1 (izquierda) y curva precision-recall con el umbral elegido (derecha), fuera de pliegue.

7. Resultados en el conjunto de prueba

Notebook: Parte II, secciones 16 a 18. Archivo: `results/evaluacion_prueba_finalistas.csv`.

Cada finalista se entrenó con todo el conjunto de desarrollo y se evaluó una sola vez sobre los 365 pacientes de prueba, con su umbral operativo:

Modelo	Umbral	Accuracy	Precisión	Recall	F1	ROC-AUC	PR-AUC
SVM RBF ajustado (reducido)	**0.26**	**0.649**	**0.481**	**0.825**	**0.607**	**0.765**	**0.588**
Random Forest ajustado (reducido)	0.34	0.660	0.490	0.808	0.610	0.759	0.589

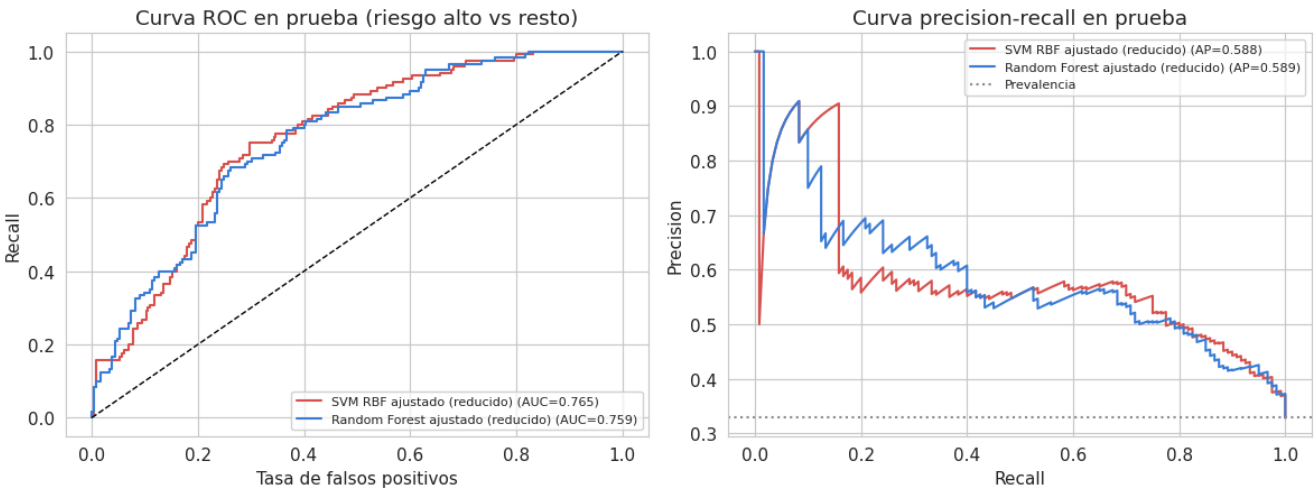


Figura 6. Curvas ROC y precision-recall de los finalistas en el conjunto de prueba.

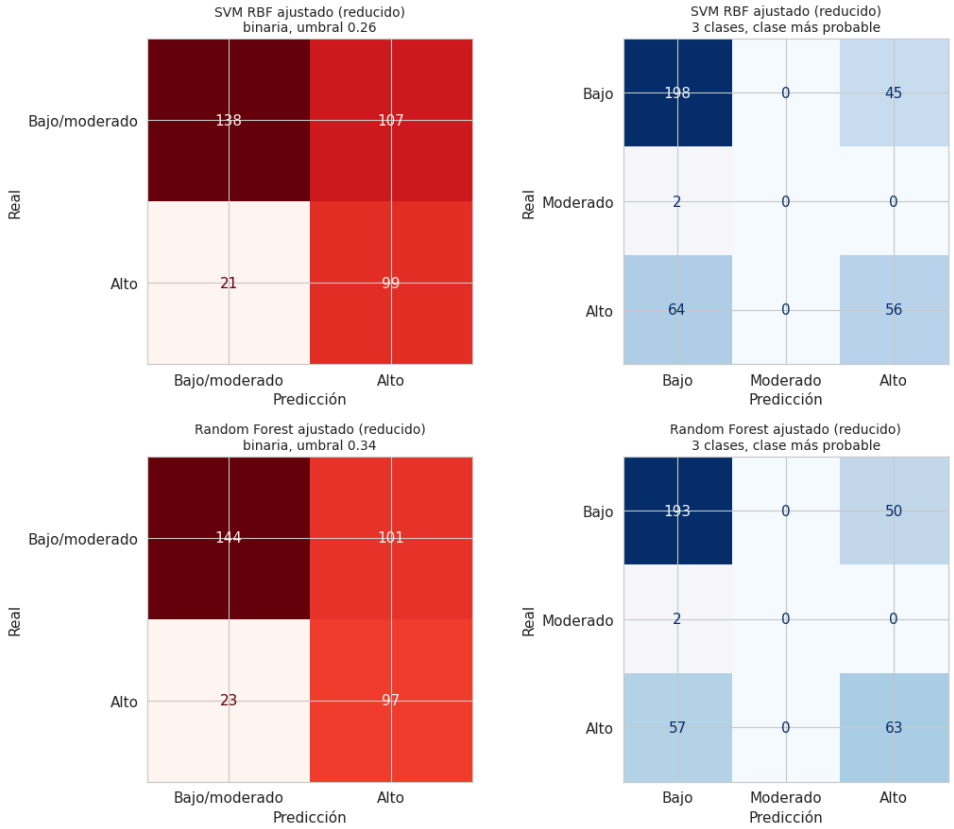


Figura 7. Matrices de confusión en prueba: decisión binaria con el umbral operativo (izquierda) y clasificación de 3 clases con la clase más probable (derecha).

Modelo final

Se eligió la **SVM RBF ajustada (reducida)**:

Criterio	Evidencia	Conclusión
Desempeño	ROC-AUC 0.764 en validación cruzada y 0.765 en prueba (Random Forest: 0.758 y 0.759)	Igual o mejor que el otro finalista y muy superior al modelo de referencia (0.505)
Generalización	ROC-AUC de prueba dentro del rango de la validación cruzada (0.764 +/- 0.027)	Sin señales de sobreajuste
Errores	Recall 0.825 con umbral 0.26: 21 falsos negativos y 107 falsos positivos	Cumple la prioridad de no perder casos de riesgo alto
Complejidad	15 variables de entrada; entrenamiento de 0.5 s	Moderada, sin variables regionales
Operación	Inferencia de unos 82 ms por cada 1,000 registros; 12 campos de entrada	Apto para inferencia por lotes

El modelo no detecta la clase moderada (12 casos en todo el dataset), por lo que se usa solo para la decisión riesgo alto frente al resto.

8. Interpretabilidad

Notebook: Parte II, secciones 19 y 20.

Se usaron dos métodos complementarios: **importancia por permutación** sobre el pipeline completo (cuánto cae el ROC-AUC de prueba al desordenar cada variable, con 20 repeticiones) y **SHAP** (KernelSHAP) sobre las variables transformadas.

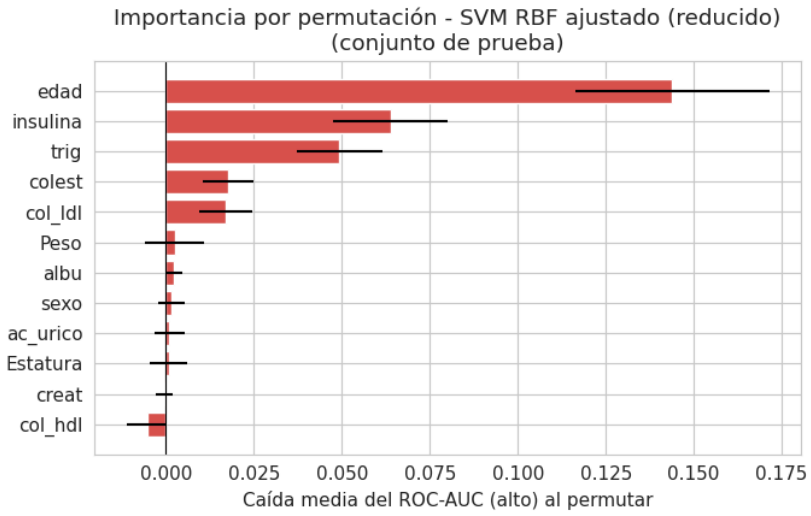


Figura 8. Importancia por permutación del modelo final en el conjunto de prueba.

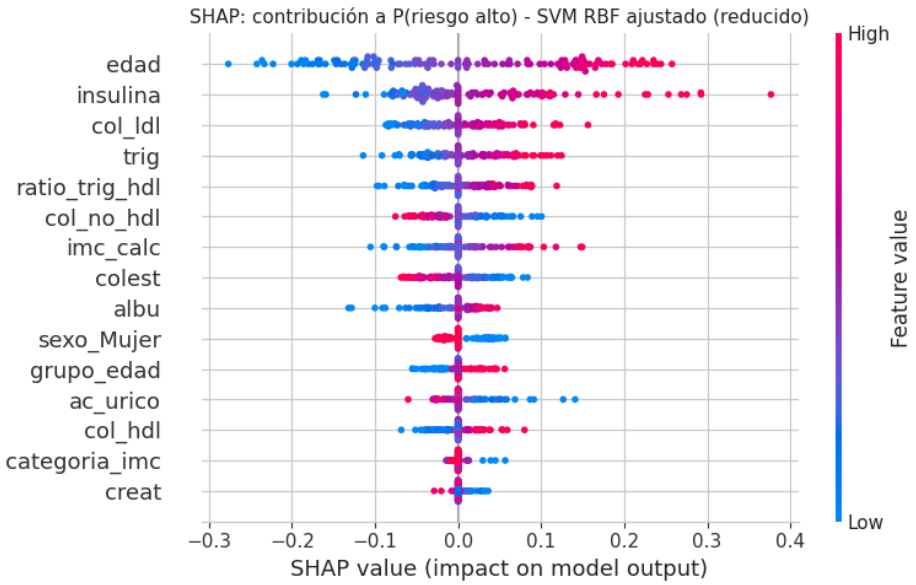


Figura 9. Contribución de cada variable a la probabilidad de riesgo alto (SHAP). Rojo: valores altos de la variable; azul: valores bajos.

La **edad** es por mucho la variable más influyente (caída de 0.144 en ROC-AUC al permutarla), seguida de la **insulina** (0.064) y los **triglicéridos** (0.049). Los valores altos de edad, insulina, triglicéridos, colesterol LDL, relación triglicéridos/HDL e IMC empujan la probabilidad hacia riesgo alto. Los valores altos de colesterol no-HDL y colesterol total la empujan ligeramente hacia abajo; como el colesterol total y el LDL están muy correlacionados (0.88), estas contribuciones deben leerse en conjunto y no por separado.

Limitaciones. La importancia por permutación subestima variables correlacionadas, porque al permutar una el modelo se compensa con la otra. KernelSHAP es una aproximación con un fondo resumido y supone independencia entre variables. Ninguno de los dos métodos implica causalidad: indican qué variables usa el modelo, no qué cambiaría el riesgo real de un paciente.

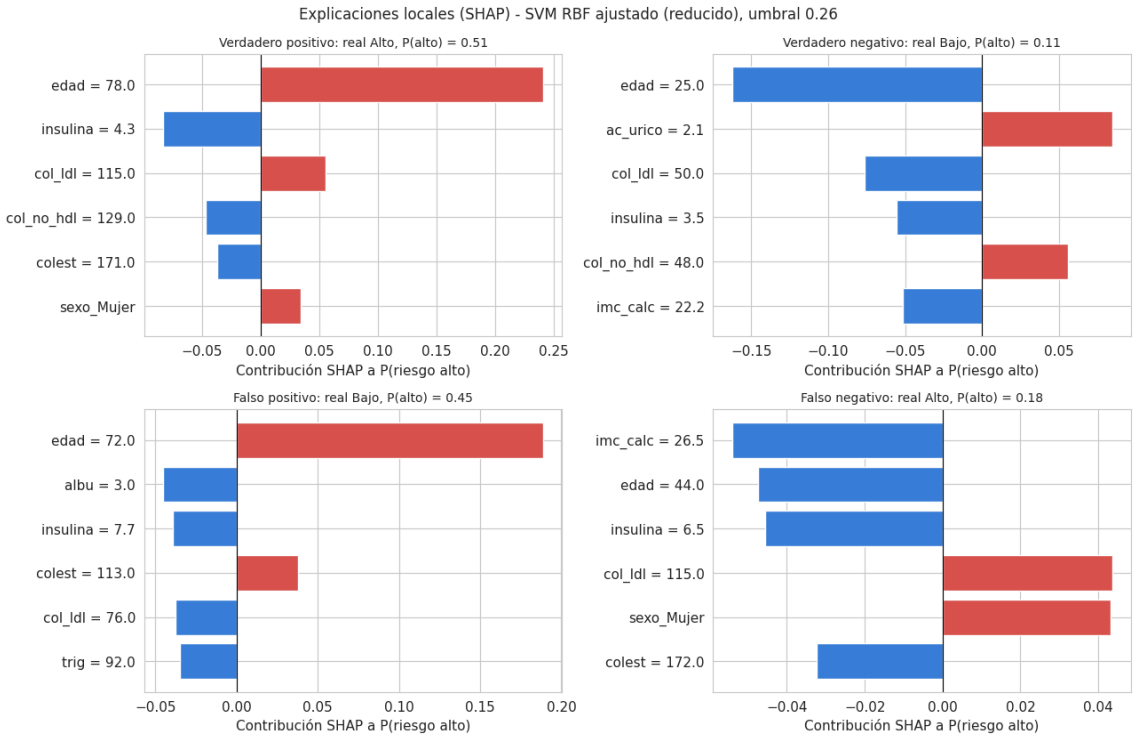


Figura 10. Explicaciones individuales de un caso representativo de cada tipo de resultado (umbral 0.26).

9. Errores, sesgos y uso responsable

Notebook: Parte II, secciones 17, 21 y 22. Archivos: `results/casos_error.csv` *y* `results/analisis_sesgos.csv`.

9.1 Análisis de errores

En prueba, con el umbral 0.26, el modelo obtiene 99 verdaderos positivos, 138 verdaderos negativos, 107 falsos positivos y 21 falsos negativos. Los falsos negativos típicos son pacientes relativamente jóvenes, con perfil lipídico y bioquímico en rango normal y, en varios casos, sin IMC medido: justo los pacientes de riesgo alto que no muestran las señales habituales. Los falsos positivos suelen tener insulina o ácido úrico elevados y mayor edad, factores que el modelo asocia correctamente con más riesgo.

Error	Impacto en este problema	Cómo reducirlo
Falso negativo	El más costoso: un paciente de riesgo alto no recibe seguimiento. Ocurre más en pacientes jóvenes sin IMC	Bajar el umbral para ese grupo o hacer obligatoria la captura de peso y estatura
Falso positivo	Una prueba confirmatoria adicional, de menor costo clínico	Aceptable con el criterio de recall >= 0.80; se comunica como prioridad, no como diagnóstico

9.2 Evaluación de sesgos

Se compararon grupos por sexo y edad con dos fuentes: el conjunto de prueba (evaluación final, grupos pequeños) y las predicciones fuera de pliegue del conjunto de desarrollo (más pacientes, más estables).

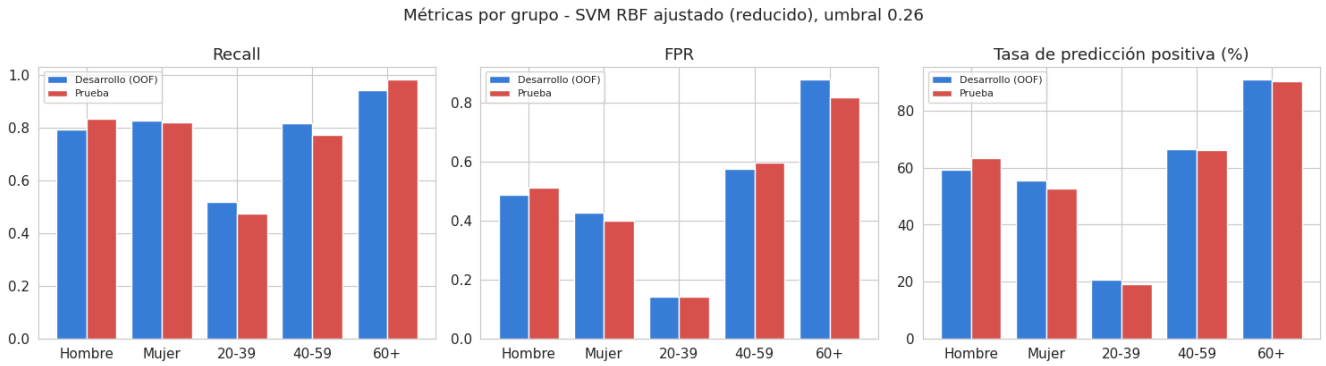


Figura 11. Recall, tasa de falsos positivos y tasa de predicción positiva por grupo.

Grupo (prueba)	Pacientes	Riesgo alto real	Recall	Precisión	FPR	FNR
Hombre	128	37.5%	0.833	0.494	0.512	0.167
Mujer	237	30.4%	0.819	0.472	0.400	0.181
20-39 años	130	14.6%	0.474	0.360	0.144	0.526
40-59 años	121	36.4%	0.773	0.425	0.597	0.227
60 años o más	111	50.5%	0.982	0.550	0.818	0.018

Por sexo, el desempeño es parejo: el recall difiere 1.4 puntos porcentuales. Por edad, la diferencia es marcada: el recall en adultos de 20 a 39 años (0.47 en prueba, 0.52 fuera de pliegue) es mucho menor que en mayores de 60 (0.98 y 0.94). En parte lo explica que el riesgo alto es mucho menos frecuente en jóvenes (14.6% frente a 50.5%) y que sus marcadores bioquímicos son menos claros. No se afirma que el modelo discrimine de forma injustificada, pero sí se documenta como limitación operativa que el personal de salud debe conocer.

9.3 Registro de riesgos

Riesgo	Causa	Consecuencia	Prob.	Impacto	Mitigación
Falsos negativos	Riesgo alto con perfil "normal"; sin IMC	Paciente en riesgo sin prueba confirmatoria	Media	Alto	Umbral con recall >= 0.80; no sustituye el tamizaje habitual
Falsos positivos	Umbral bajo para favorecer el recall	Pruebas adicionales y preocupación	Alta	Bajo a medio	Comunicar como "prioridad de revisión"
Deriva de datos	Cambios en la población o en el laboratorio	Probabilidades descalibradas	Media	Alto	Monitoreo mensual de PSI y recall
Datos incompletos	Peso y estatura faltantes (28%)	Imputación menos confiable	Alta	Medio	Validación de entradas con advertencias
Uso automático	Tomar la clasificación como decisión final	Negar o retrasar atención	Media	Alto	Revisión humana obligatoria
Sesgo por grupo	Diferente prevalencia y calidad de datos	Menor recall en jóvenes	Media	Medio a alto	Reporte trimestral por grupo; revisión si la brecha supera 0.10

9.4 Condiciones de uso

- **Usos permitidos:** priorizar pruebas confirmatorias y acciones preventivas en una población que ya cuenta con perfil de lípidos; apoyar la planeación de campañas.
- **Usos no recomendados:** diagnosticar diabetes; negar, retrasar o condicionar atención, seguros o empleo; aplicarlo a menores de 20 años, embarazadas o poblaciones distintas a la de la encuesta; estimar el riesgo moderado.
- **Revisión humana:** todo resultado lo revisa personal de salud; los casos cercanos al umbral (+/- 0.05) y los registros con advertencias reciben revisión adicional.
- **Suspensión del modelo:** recall de riesgo alto menor a 0.70 en dos cortes mensuales seguidos, PSI mayor a 0.25 en una variable principal, brecha de recall entre grupos mayor a 0.15, cambios en la definición del objetivo o falla en la verificación del artefacto.

10. Puesta en operación

Notebook: Parte II, secciones 23 a 29.

10.1 Artefactos

El pipeline completo se serializa con Joblib junto con la codificación del objetivo y el umbral. Cada artefacto se registra con su versión y su huella SHA-256 para verificar su integridad antes de cargarlo.

Artefacto	Archivo	Tamaño
Pipeline final	`artifacts/pipeline_riesgo_diabetes_v1.joblib`	258.9 KB
Clasificador	`artifacts/modelo_clasificador_v1.joblib`	252.1 KB
Codificación del objetivo	`artifacts/codificacion_objetivo_v1.json`	0.1 KB
Umbral	`artifacts/umbral_v1.json`	0.1 KB
Metadatos y esquema de entrada	`artifacts/metadatos_modelo_v1.json`	4.1 KB

> **Seguridad:** los archivos .joblib pueden ejecutar código al cargarse. Nunca deben cargarse artefactos de fuentes no verificadas; antes de cada carga se compara su huella SHA-256 con la registrada.

10.2 Inferencia con validación de entradas

El script independiente `scripts/inferencia_riesgo_diabetes.py` carga el artefacto, recibe registros nuevos (diccionario o CSV), valida su estructura, calcula la probabilidad de riesgo alto, aplica el umbral y devuelve la clasificación ("prioridad alta" o "sin prioridad alta") con un recordatorio de revisión humana. Sus predicciones coinciden exactamente con las del modelo antes de serializarse en los 365 registros de prueba.

Caso de prueba	Entrada	Resultado esperado	Resultado obtenido
Registro válido	Registro completo	Válido	Válido con advertencia (columna extra `glu_suero` ignorada)
Campo obligatorio faltante	Sin `trig`	Rechazado	Rechazado: campo faltante "trig"
Campo opcional faltante	Sin peso	Válido con advertencia	Válido con advertencia: se imputa con la mediana
Categoría desconocida	sexo = "Otro"	Rechazado	Rechazado: categoría no reconocida
Tipo incorrecto	edad = "cuarenta"	Rechazado	Rechazado: tipo incorrecto en "edad"

10.3 Simulación con pacientes nuevos

Perfil	P(riesgo alto)	Clasificación	Acción sugerida
Riesgo bajo	0.061	Sin prioridad alta	Seguimiento habitual y consejo preventivo
Riesgo medio	0.277	Prioridad alta	Revisar en consulta; el personal decide si adelanta la prueba
Riesgo alto	0.697	Prioridad alta	Priorizar prueba confirmatoria y consejería
Combinación poco frecuente (hombre, 60+, IMC < 25)	0.370	Prioridad alta	Estimación menos respaldada por datos: revisión clínica completa
Caso real cercano al umbral	0.260	Prioridad alta	Caso frontera: decide el personal de salud

10.4 Plan de monitoreo

Indicador	Frecuencia	Alerta	Acción
Valores faltantes por campo	Semanal	+10 puntos sobre lo observado en entrenamiento	Revisar la captura
Registros rechazados por validación	Semanal	Más de 5%	Revisar formatos y unidades
Deriva de datos (PSI) en edad, IMC, triglicéridos, HDL e insulina	Mensual	PSI > 0.10 aviso; > 0.25 alerta	Analizar causa; recalibrar o reentrenar
Recall de riesgo alto	Mensual	< 0.75 aviso; < 0.70 en dos cortes: suspender	Ajustar umbral o reentrenar
Brecha de recall entre grupos	Trimestral	> 0.10 aviso; > 0.15 suspender	Análisis de sesgo
Tiempo de respuesta	Continuo	Más de 2 s por lote de 1,000	Revisar infraestructura

Se simuló el cálculo del PSI: entre desarrollo y prueba la población es estable (PSI de edad = 0.021). Con un lote simulado con deriva deliberada (triglicéridos 30% más altos, 35% de faltantes en peso y estatura y población de mayor edad), el PSI de edad sube a 0.827 y dispara correctamente la alerta.

11. Reproducibilidad

Pasos para reproducir todos los resultados:

1. Clonar el repositorio github.com/AlanDelAngel/diabetes_pacientes_mexicanos.
2. Descargar el dataset desde Kaggle y guardar `Diabetes_Mexico.csv` junto al notebook o en `data/`.
3. Instalar dependencias con `pip install -r requirements.txt`.
4. Ejecutar `Proyecto_Final_Riesgo_Diabetes.ipynb` de principio a fin (también funciona en Google Colab).

Carpeta generada	Contenido
`scripts/`	Transformador propio del pipeline y script de inferencia
`artifacts/`	Pipeline serializado, clasificador, umbral, codificación y metadatos
`results/`	Registro de experimentos, comparación de modelos, evaluación en prueba, sesgos, errores y figuras
`docs/`	Ficha del modelo en Markdown

Elemento	Detalle
Versión del dataset	`Diabetes_Mexico.csv`, 2,549 registros x 24 columnas, md5 `eb696bc3e9a018f03290f6265e7cabfe`
Partición	StratifiedGroupKFold por hogar, semilla 42: desarrollo 2,184 / prueba 365
Experimentos	19 configuraciones registradas con fecha, huella del dataset, pipeline, hiperparámetros y métricas
Bibliotecas	Python 3.11, scikit-learn 1.8.0, pandas 3.0.2, NumPy 2.4.4, Joblib 1.5.3, SHAP 0.51.0
Fecha de entrenamiento	30 de septiembre de 2026

12. Ficha del modelo

Campo	Contenido
Nombre y versión	Modelo de priorización de riesgo de diabetes, v1.0.0 (SVM RBF ajustado, modelo reducido)
Propósito	Ordenar a qué pacientes adultos conviene hacer primero una prueba confirmatoria de diabetes, con datos demográficos, antropométricos y un perfil bioquímico sin glucosa ni HbA1c
Usuarios previstos	Personal de epidemiología y planeación en salud, médicos de primer contacto y coordinadores de programas preventivos
Dataset	2,549 adultos de 2,055 hogares; desarrollo 2,184 y prueba 365
Hiperparámetros	Kernel RBF, C = 3, gamma = 0.03; 15 variables de entrada
Métricas	Validación cruzada: ROC-AUC 0.764 +/- 0.027. Prueba: accuracy 0.649, precisión 0.481, recall 0.825, F1 0.607, ROC-AUC 0.765, PR-AUC 0.588
Umbral	0.26 sobre la probabilidad de riesgo alto, elegido para recall >= 0.80
Limitaciones	Desempeño moderado (ROC-AUC cercano a 0.76); no detecta la clase moderada; requiere perfil de lípidos e insulina; 28% sin peso o estatura; datos de una sola encuesta; menor recall en 20-39 años
Análisis de sesgos	Recall en prueba: hombres 0.83, mujeres 0.82; 20-39 años 0.47; 40-59 años 0.77; 60+ años 0.98
Monitoreo	Calidad de datos semanal; PSI y porcentaje de señalados mensual; recall, FPR, FNR y equidad trimestral
Actualización	Recall < 0.75 en dos cortes, PSI > 0.25 o brecha de recall entre grupos > 0.15; revisión anual
Responsables	Alan Adelfo Del Angel Figueroa (técnico) y coordinación de epidemiología (clínico)

Aprendizajes

Este proyecto me permitió recorrer de principio a fin el ciclo de vida de un modelo aplicado a un problema real de salud pública: desde detectar errores en los propios datos (el significado de la variable objetivo, un valor centinela que se colaba al modelo) hasta definir cómo debería monitorearse una vez en uso.

Lo que más me hizo reflexionar fue el análisis de sesgos. El modelo es parejo por sexo, pero mucho más débil en adultos jóvenes, simplemente porque el riesgo alto es raro en ese grupo y hay menos señal para aprender. Una métrica global como un ROC-AUC de 0.76 puede esconder debilidades importantes en subgrupos, y reportarlas con transparencia es parte central de construir un modelo responsable. También aprendí que elegir el umbral con un criterio clínico explícito es una decisión tan importante como elegir el modelo, y que un modelo de este tipo nunca debería usarse para negar atención o diagnosticar, sino para apoyar la decisión de un profesional de salud.

Referencias

American Diabetes Association. (2023). Classification and diagnosis of diabetes: Standards of Care in Diabetes 2023. *Diabetes Care, 46*(Supplement 1), S19-S40. https://doi.org/10.2337/dc23-S002

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5-32. https://doi.org/10.1023/A:1010933404324

Félix Jiménez, A. F. y Sánchez Lee, V. S. (2026). *Diabetes Mexico Data Set* [Conjunto de datos]. Kaggle. https://doi.org/10.34740/KAGGLE/DSV/14619160

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

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825-2830.