IDENTIFICACIÓN DE CARACTERÍSTICAS CLAVE PARA LA VALORACIÓN DE RIESGO DE DIABETES Y CLASIFICACIÓN DE PACIENTES

Samsung
UDEM

Integrantes:
Moisés de Jesús Juárez Álvarez
Marcos Emmanuel Mariscal Díaz
Paul Ivan Ramos de la Torre
Profesores:
Mtro. Eduardo de Ávila Armenta
Dr. Alberto Luque Chang

ÍNDICE

INTRODUCCIÓN

La diabetes tipo 2 (DM2) es el 90% de los casos de diabetes, con 700 millones de proyectados para 2045. Factores clave: obesidad, envejecimiento y sedentarismo.

Limitaciones del diagnóstico actual:

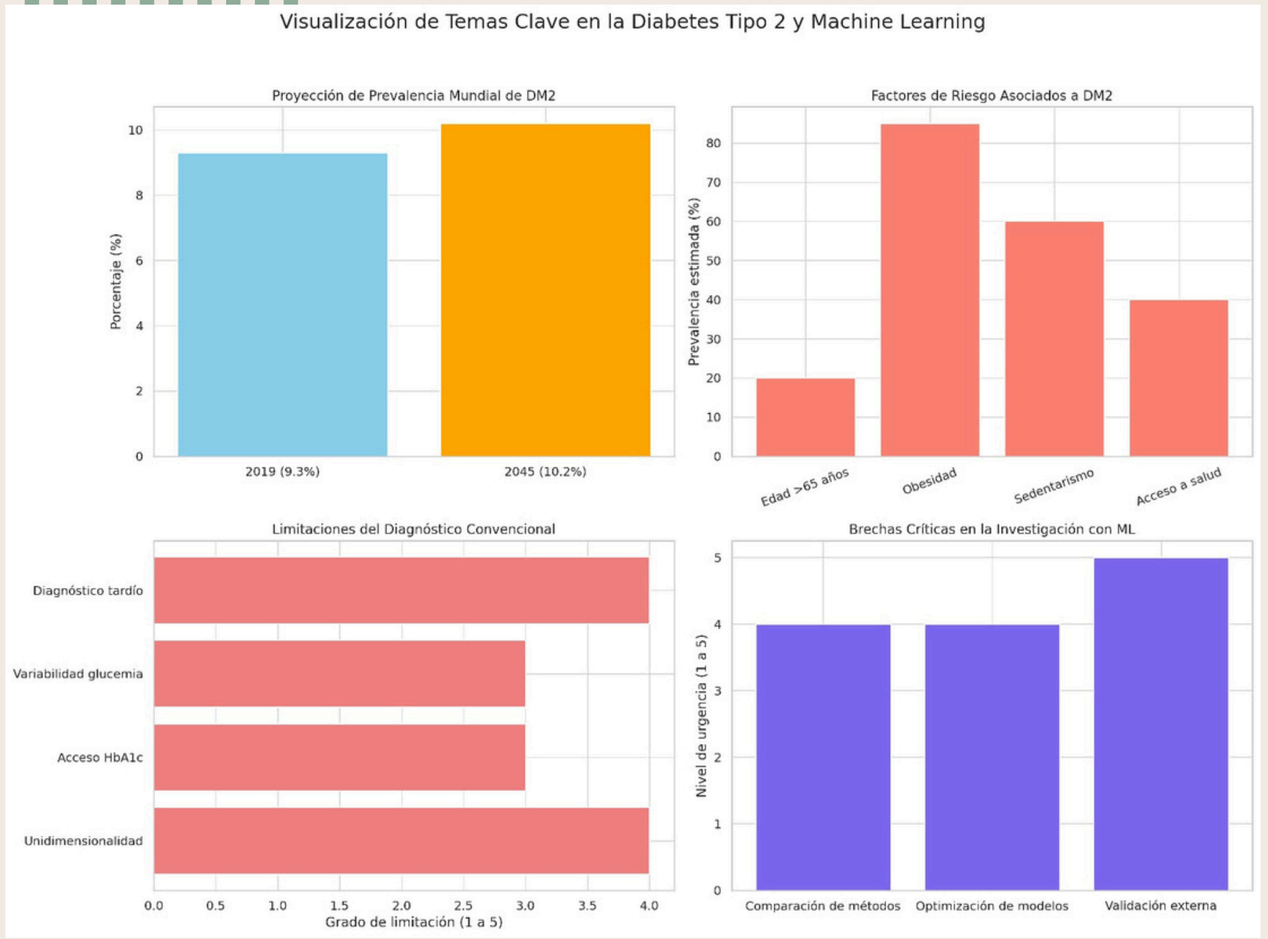
- Detección tardía (cuando ya hay complicaciones).
- Pruebas como glucosa o HbA1c son variables o poco accesibles.

Machine Learning (ML) como solución:

- Modelos predictivos logran 70–95% de precisión.
- Técnicas como Random Forest superan a métodos tradicionales.

Brechas del estudio:

1. Falta comparación de métodos para seleccionar variables clave.
2. Optimización insuficiente de modelos.
3. Poca validación externa.



[1] International Diabetes Federation. IDF Diabetes Atlas, 9th edition. Brussels, Belgium: IDF; 2019. Disponible en: https://www.diabetesatlas.org [2] World Health Organization. Global report on diabetes. Geneva: WHO; 2020. ISBN: 978-92-4-156525-7 [3] American Diabetes Association. 2. Classification and Diagnosis of Diabetes: Standards of Medical Care in Diabetes—2021. Diabetes Care. 2021;44(Supplement_1):S15-S33. DOI: 10.2337/dc21-S002

OBJETIVOS

General

- Este estudio desarrolla un modelo predictivo de diabetes tipo 2 mediante machine learning y selección de características, superando las limitaciones del diagnóstico convencional. Busca integrar variables clínicas, demográficas y de estilo de vida para mejorar la precisión por encima del 90 %, reduciendo la cantidad de variables a menos de 10 para facilitar su implementación en entornos clínicos con recursos limitados.

Área	Aporte esperado
Clínica	Protocolo de screening simplificado con 7-10 variables clave
Tecnológica	Pipeline reproducible para selección de características médicas
Metodológica	Framework comparativo Lasso vs. Boruta en datos clínicos reales
Social	Reducción potencial de costos diagnósticos en poblaciones marginadas

Específicos

- Objetivo del estudio:
- Analizar una población mediante 46 variables clínicas y demográficas.
- Evaluación estadística:
- Determinar el poder predictivo de las variables utilizando:
- Pruebas estadísticas.
- Comparación de técnicas de selección de características:
- Lasso (Regresión con penalización L1).
- Boruta (algoritmo basado en Random Forest).
- Optimización de modelos de machine learning:
- Random Forest (bosques aleatorios).
- Regresión Logística.
- SVM (Máquinas de Vectores de Soporte).
- Interpretación clínica:
- Identificación de variables clave y su relevancia predictiva.
- Relación de los hallazgos con guías médicas actuales.

HIPÓTESIS

- H1: La combinación de variables clínicas (HbA1c) y síntomas clásicos (poliuria) tendrá mayor poder predictivo que los marcadores metabólicos aislados
- H2: Los modelos con selección de características superarán en generalización a los que usan todas las variables
- H3: Las variables socioeconómicas mostrarán interacciones significativas con marcadores clínicos

METODOLOGÍA

Etapa 1

1

Este estudio emplea un diseño analítico observacional con enfoque predictivo, basado en el análisis secundario de un conjunto de datos clínicos retrospectivos.

Etapa 2

2

El conjunto de datos incluye registros de 1,879 pacientes adultos (20-90 años) con las siguientes características:

Variable	Valor	Fuente
Edad (años)	58.7 ± 12.3	Diabetes.csv
Género (% femenino)	48.7 %	Diabetes.csv
Prevalencia diabetes	40 %	Diabetes.csv
Número de características	46	Diabetes.csv

Etapa 3

3

1. Las variables analizadas se agrupan en ocho categorías: datos demográficos, estilo de vida, historial médico, mediciones clínicas, uso de medicamentos, síntomas y calidad de vida, exposición ambiental y comportamientos de salud. Estas incluyen factores como edad, IMC, presión arterial, hábitos de salud y exposición a riesgos ambientales, permitiendo una evaluación integral del riesgo de diabetes tipo 2.

Etapa 4

4

Se aplicaron las siguientes transformaciones: Manejo de valores faltantes: No fue necesario ya que el dataset estaba sin valores faltantes. División de datos: Manejo adecuado de datos. Balanceo: Estratificación para preservar proporción de clases

Etapa 5

5

Selección de Características

- Lasso: Regularización L1 con α = 0,02 (óptimo via validación cruzada)
- Boruta: 100 iteraciones con Random Forest (profundidad=5

Etapa 6

6

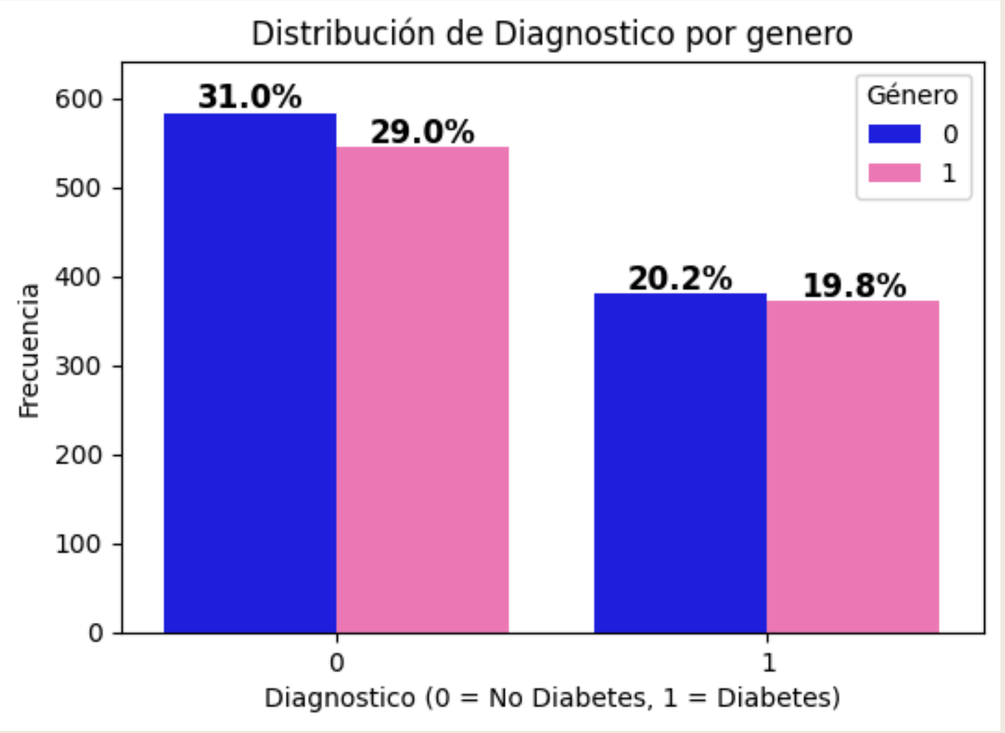
Las métricas utilizadas para evaluar el modelo incluyen exactitud (accuracy), precisión, sensibilidad, F1-score. Estas permiten medir el rendimiento del modelo en la identificación de pacientes con diabetes tipo 2, equilibrando la detección correcta y la minimización de errores.

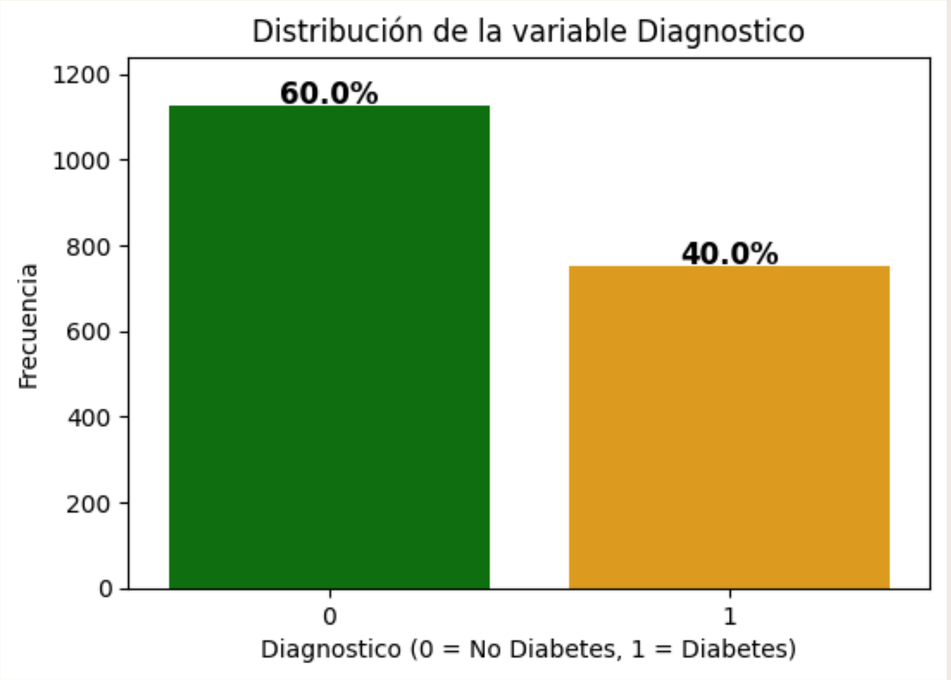
[4] Kavakiotis I, Tsave O, Salifoglou A, et al. Machine learning and data mining methods in diabetes research. Comput Struct Biotechnol J. 2017;15:104-116. DOI: 10.1016/j.csbj.2016.12.005 [5] Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. BMJ. 2015;350:g7594. DOI: 10.1136/bmj.g7594

RESULTADOS

El análisis inicial reveló una distribución heterogénea en las variables clave:

Variable	No Diabetes (n=1,127)	Diabetes (n=752)	p-valor
Edad (años)	54.3 ± 11.2	65.1 ± 10.8	<0.001
IMC (kg/m²)	26.8 ± 4.9	30.7 ± 6.1	<0.001
HbA1c (%)	5.4 ± 0.4	7.2 ± 1.3	<0.001
Presión Sistólica (mmHg)	122.3 ± 12.1	138.4 ± 15.8	<0.001

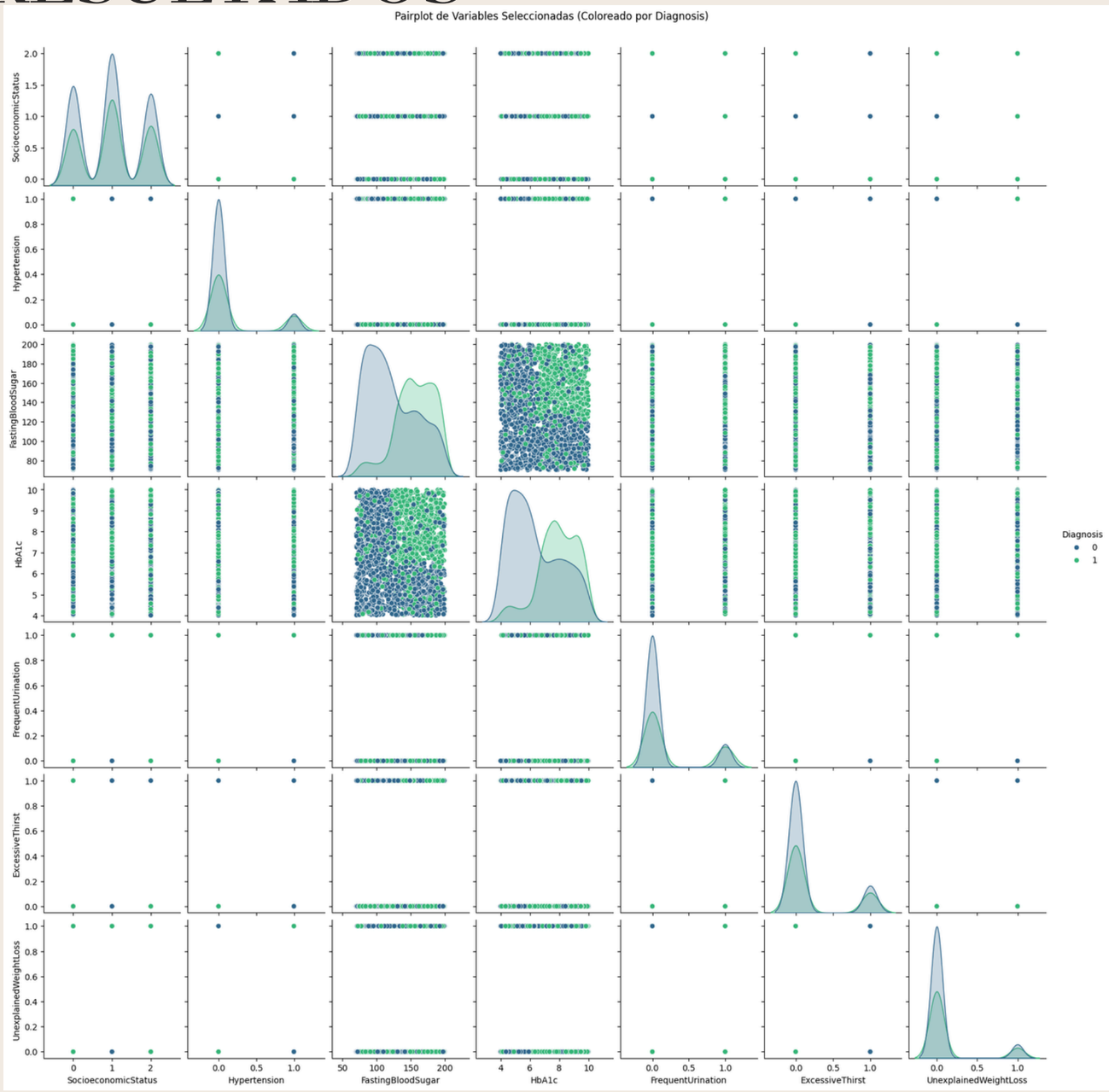




RESULTADOS

Lasso identificó 7 predictores clave:

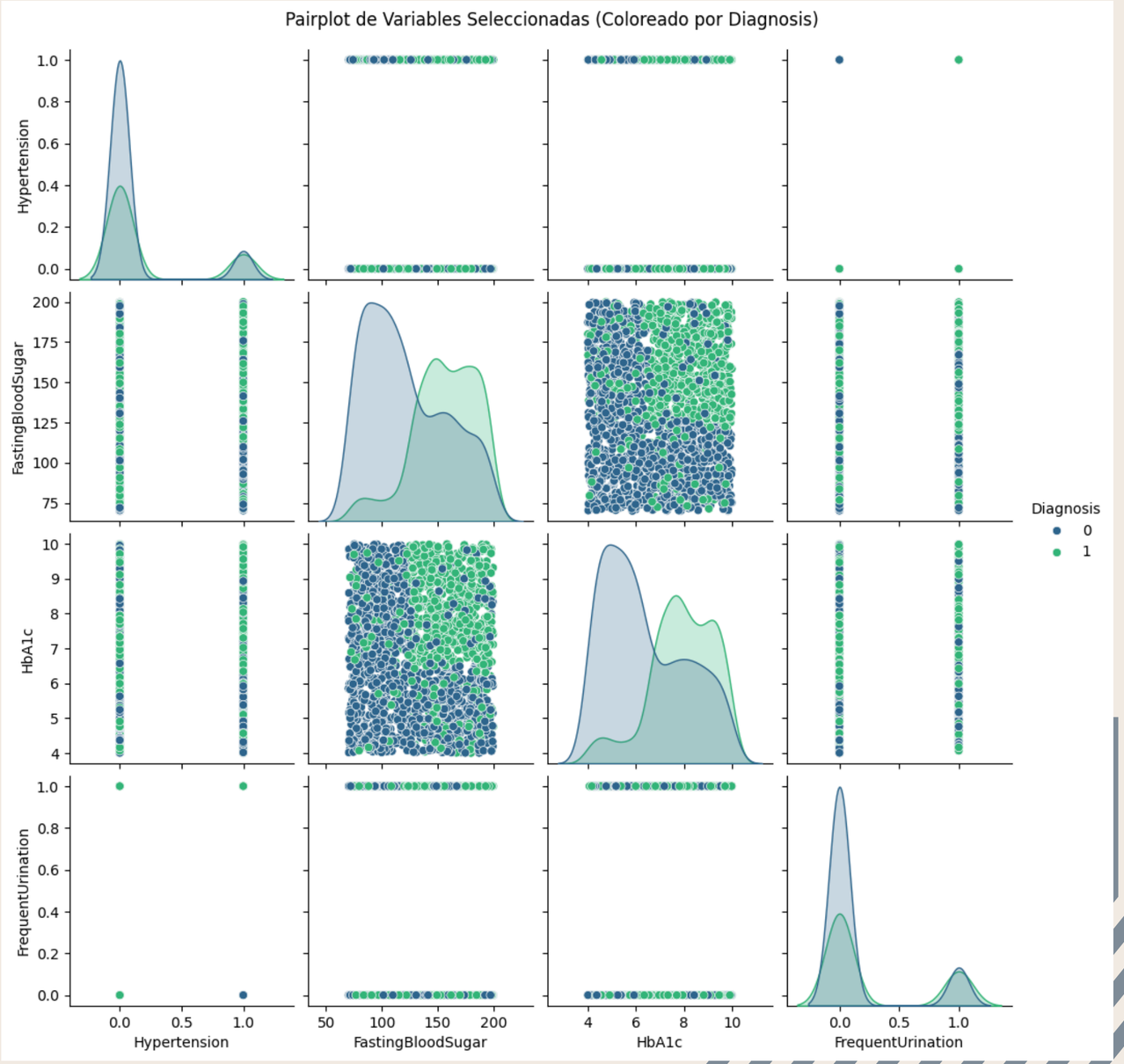
1. Estatus socioeconómico (SocioeconomicStatus)
2. Hipertensión (Hypertension)
3. Azúcar en ayunas (FastingBloodSugar)
4. HbA1c
5. Poliuria (FrequentUrination)
6. Sed excesiva (ExcessiveThirst)
7. Pérdida de peso inexplicable (UnexplainedWeightLoss)



RESULTADOS

Borutta confirmó 4 características relevantes:

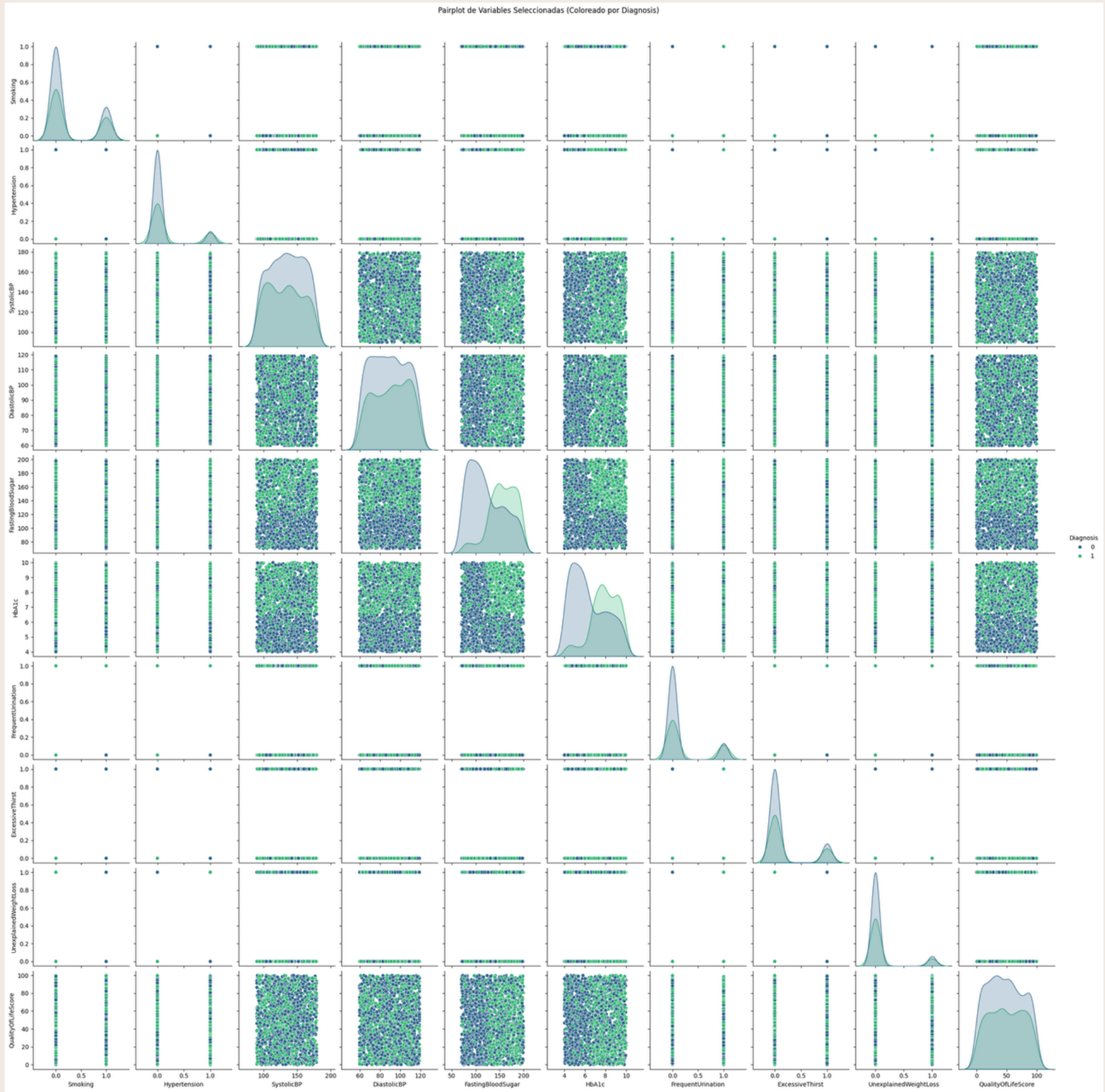
1. Hipertensión (Hypertension)
2. Azúcar en ayunas (FastingBloodSugar)
3. HbA1c
4. Poliuria (FrequentUrination)



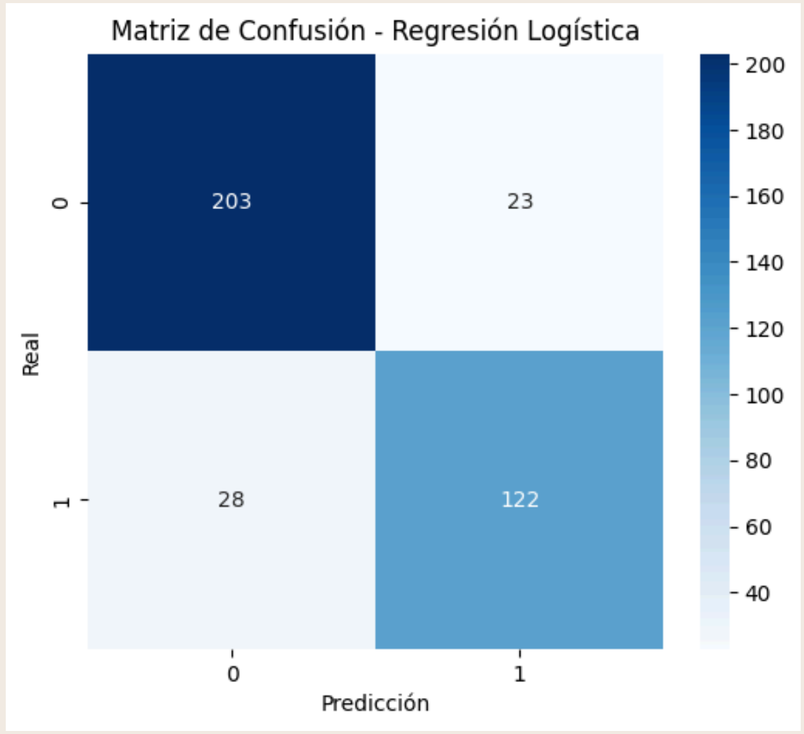
RESULTADOS

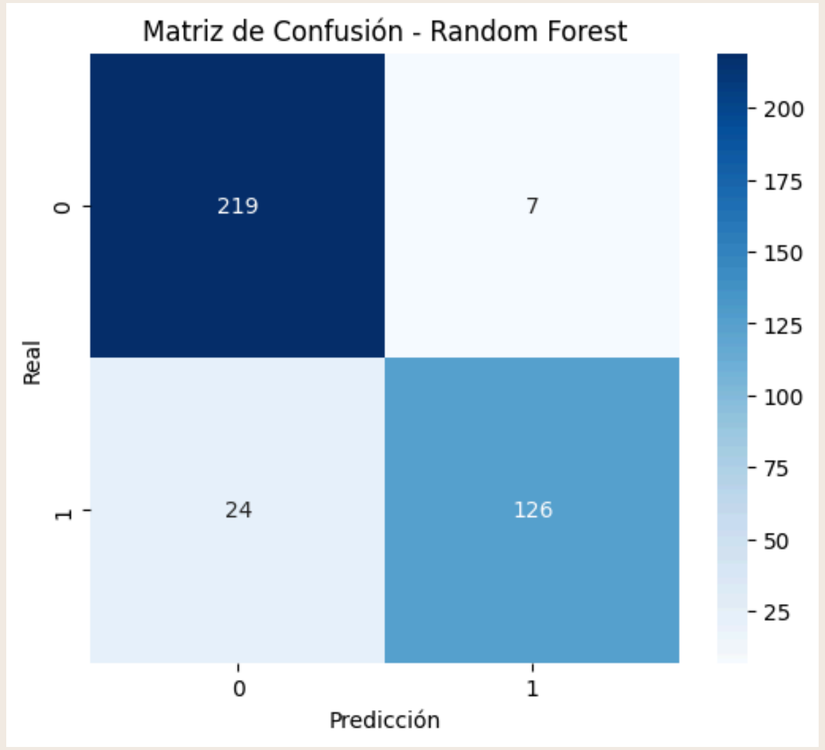
El algoritmo Selectkbest confirmó 10 características relevantes:

1. Fumar (Smoking)
2. Hipertensión (Hypertension)
3. Presión arterial sistólica (SystolicBP)
4. Presión arterial diastólica (DiastolicBP)
5. Azúcar en ayunas (FastingBloodSugar)
6. HbA1c
7. Poliuria (FrequentUrination)
8. Sed excesiva (ExcessiveThirst)
9. Pérdida de peso inexplicable (UnexplainedWeightLoss)
10. Puntuación de calidad de vida (QualityOfLifeScore)



MATRICES DE CONFUSION





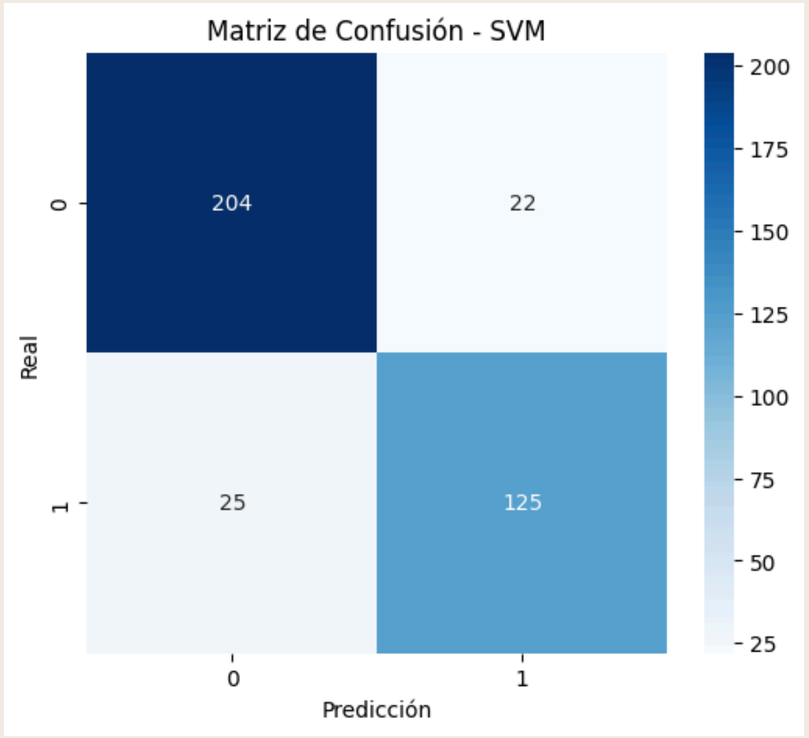


Table 10: Métricas de desempeño del modelo Random Forest

Clase	Precisión	Sensibilidad	F1-score	Soporte
Clase 0	0.90	0.97	0.93	226
Clase 1	0.95	0.84	0.89	150
Promedio macro	0.92	0.90	0.91	–
Promedio ponderado	0.92	0.92	0.92	–
Precisión global (ACC): 0.92				

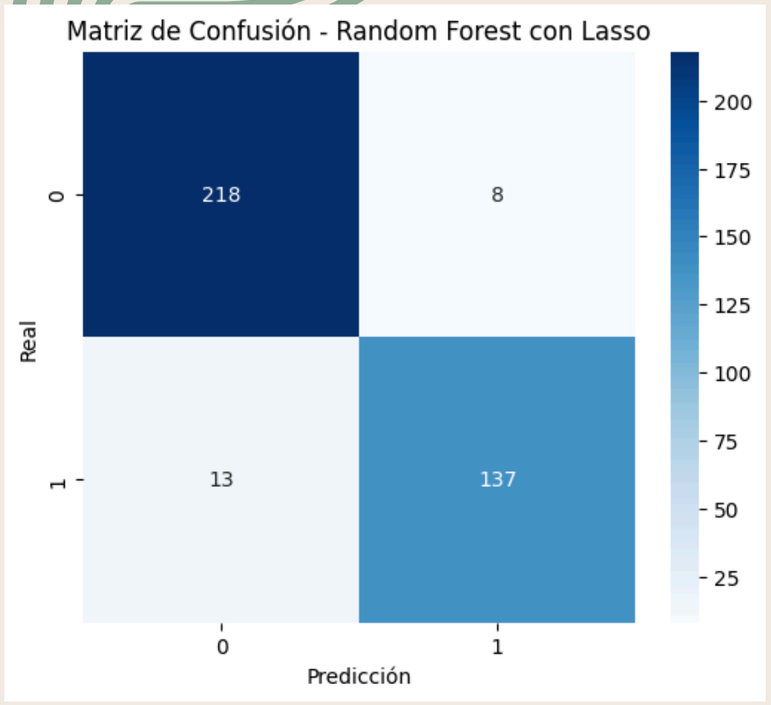
Table 11: Métricas de desempeño del modelo Regresion Logistica

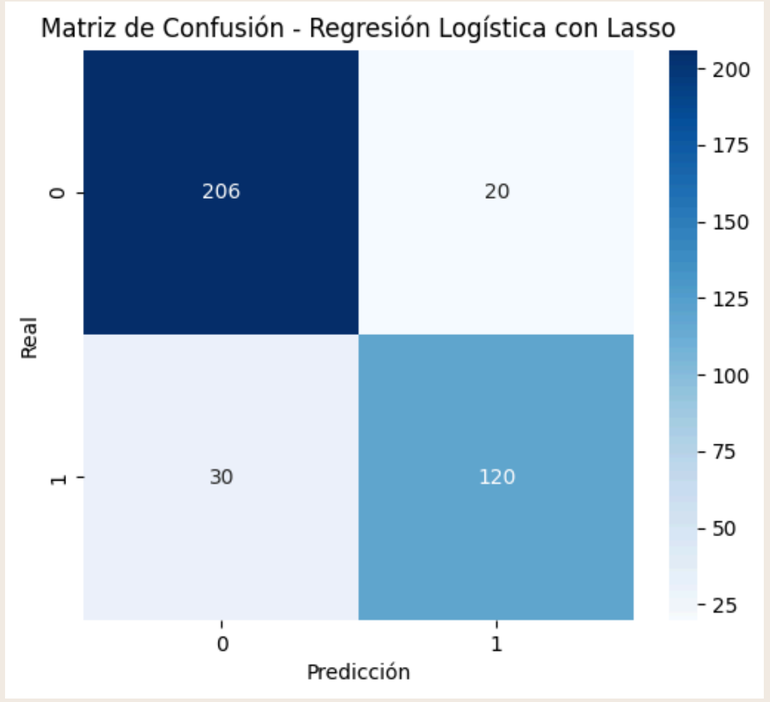
Clase	Precisión	Sensibilidad	F1-score	Soporte
Clase 0	0.88	0.90	0.89	226
Clase 1	0.84	0.81	0.83	150
Promedio macro	0.86	0.86	0.86	–
Promedio ponderado	0.86	0.86	0.86	–
Precisión global (ACC): 0.86				

Table 12: Métricas de desempeño del modelo SVM

Clase	Precisión	Sensibilidad	F1-score	Soporte
Clase 0	0.89	0.90	0.90	226
Clase 1	0.85	0.83	0.84	150
Promedio macro	0.87	0.87	0.87	–
Promedio ponderado	0.87	0.88	0.87	–
Precisión global (ACC): 0.88				

MATRICES DE CONFUSION LASSO





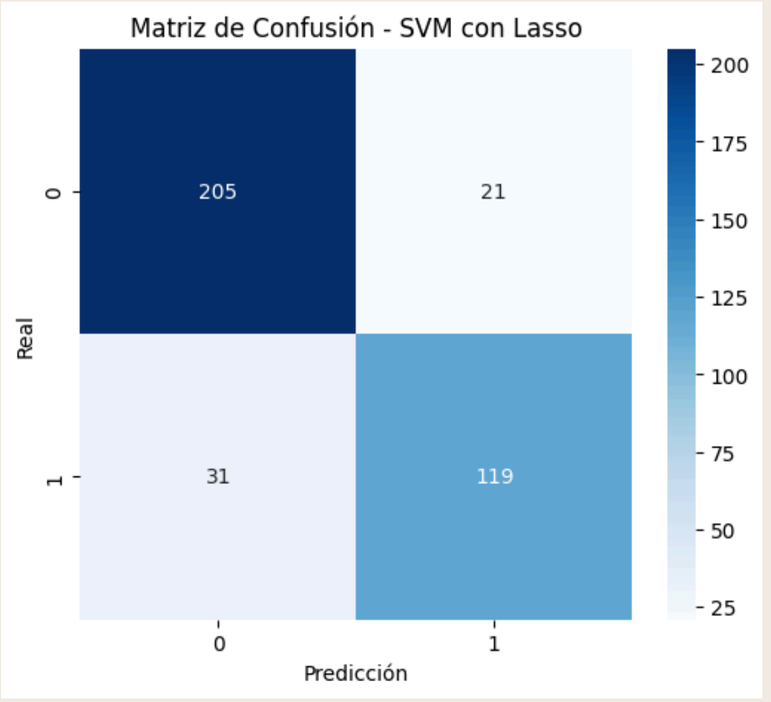


Table 4: Métricas de desempeño del modelo Random Forest con Lasso

Clase	Precisión	Sensibilidad	F1-score	Soporte
Clase 0	0.94	0.96	0.95	226
Clase 1	0.94	0.91	0.93	150
Promedio macro	0.94	0.94	0.94	–
Promedio ponderado	0.94	0.94	0.94	–
Precisión global (ACC): 0.94				

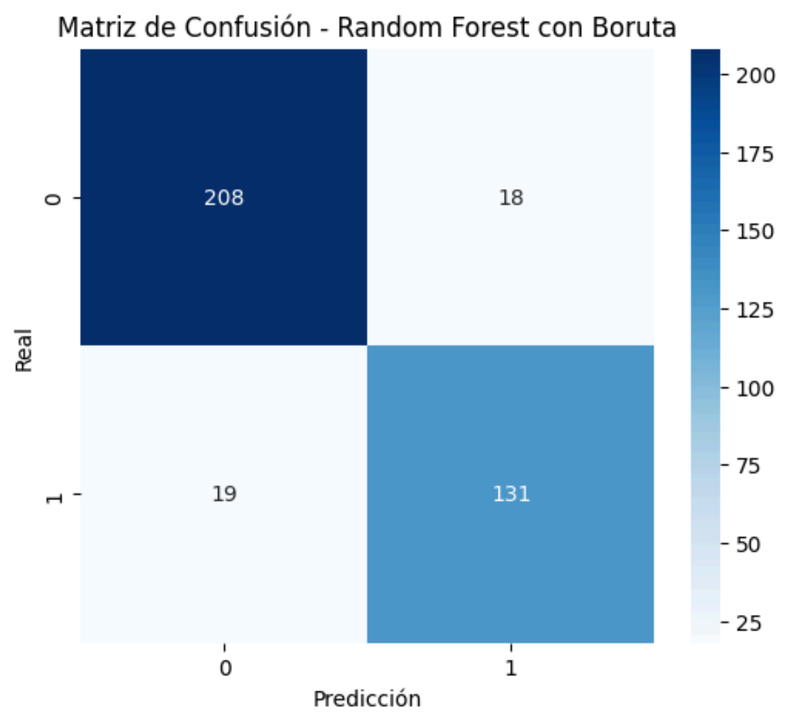
Table 5: Métricas de desempeño del modelo Regresion logistica con Lasso

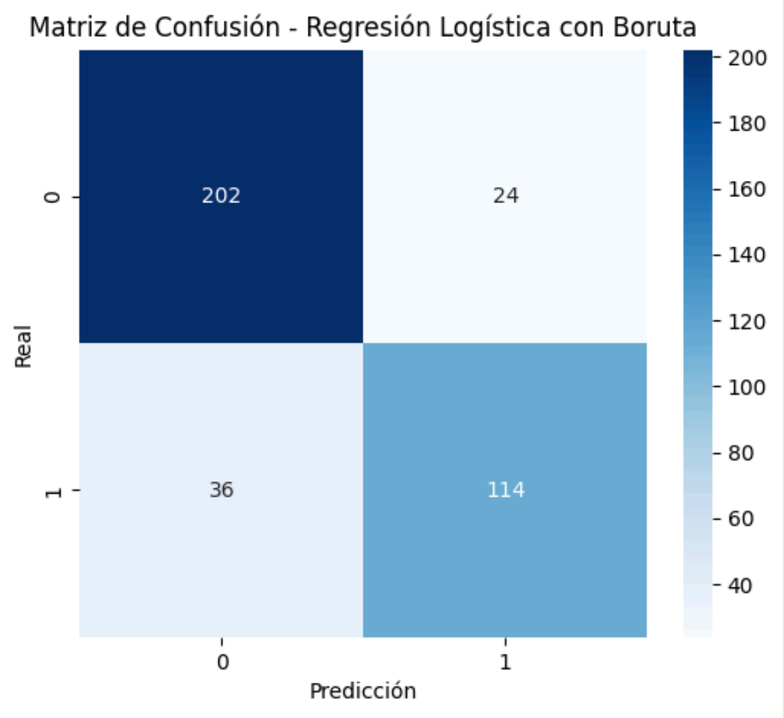
Clase	Precisión	Sensibilidad	F1-score	Soporte
Clase 0	0.87	0.91	0.89	226
Clase 1	0.86	0.80	0.83	150
Promedio macro	0.87	0.86	0.86	–
Promedio ponderado	0.87	0.87	0.87	–
Precisión global (ACC): 0.87				

Table 6: Métricas de desempeño del modelo SVM con Lasso

Clase	Precisión	Sensibilidad	F1-score	Soporte
Clase 0	0.87	0.91	0.89	226
Clase 1	0.85	0.79	0.82	150
Promedio macro	0.86	0.85	0.85	–
Promedio ponderado	0.86	0.86	0.86	–
Precisión global (ACC): 0.86				

MATRICES DE CONFUSION BORUTA





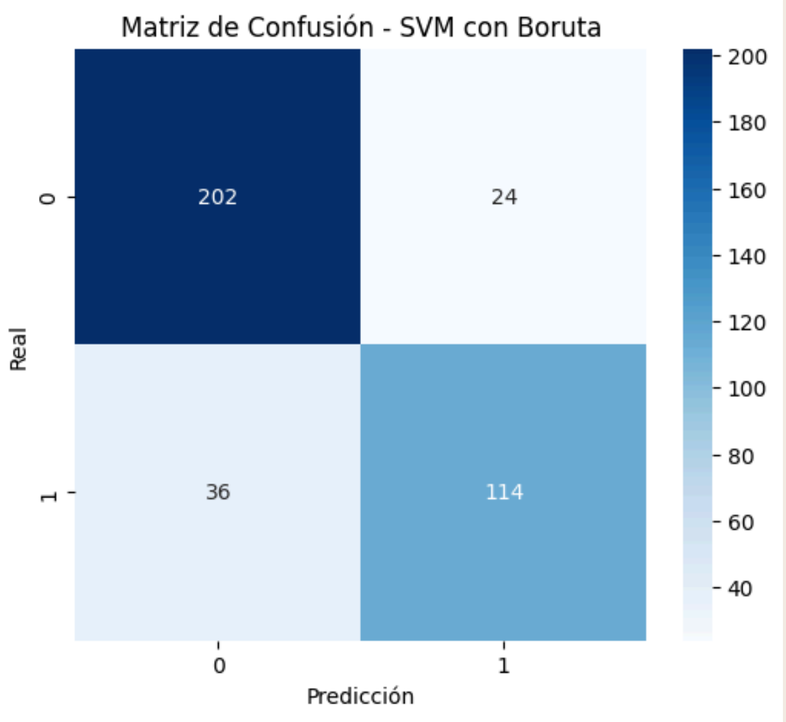


Table 7: Métricas de desempeño del modelo Random Forest con Boruta

Clase	Precisión	Sensibilidad	F1-score	Soporte
Clase 0	0.92	0.92	0.92	226
Clase 1	0.88	0.87	0.88	150
Promedio macro	0.90	0.90	0.90	–
Promedio ponderado	0.90	0.90	0.90	–
Precisión global (ACC): 0.90				

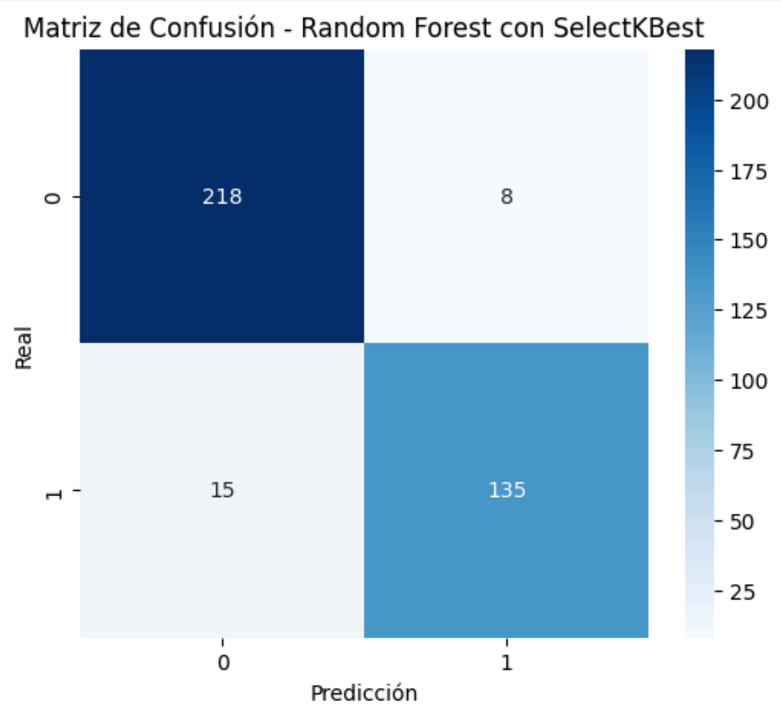
Table 8: Métricas de desempeño del modelo Regresion Logistica con Boruta

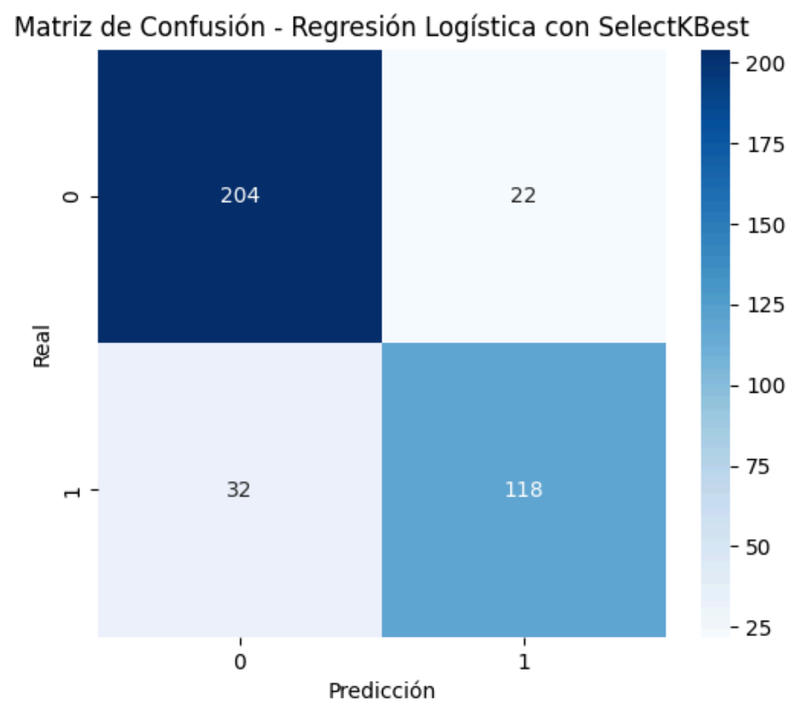
Clase	Precisión	Sensibilidad	F1-score	Soporte
Clase 0	0.85	0.89	0.87	226
Clase 1	0.83	0.76	0.79	150
Promedio macro	0.83	0.83	0.83	–
Promedio ponderado	0.84	0.84	0.84	–
Precisión global (ACC): 0.84				

Table 9: Métricas de desempeño del modelo SVM con Boruta

Clase	Precisión	Sensibilidad	F1-score	Soporte
Clase 0	0.85	0.89	0.87	226
Clase 1	0.83	0.76	0.79	150
Promedio macro	0.83	0.83	0.83	–
Promedio ponderado	0.84	0.84	0.84	–
Precisión global (ACC): 0.84				

MATRICES DE CONFUSION SELECKBEST





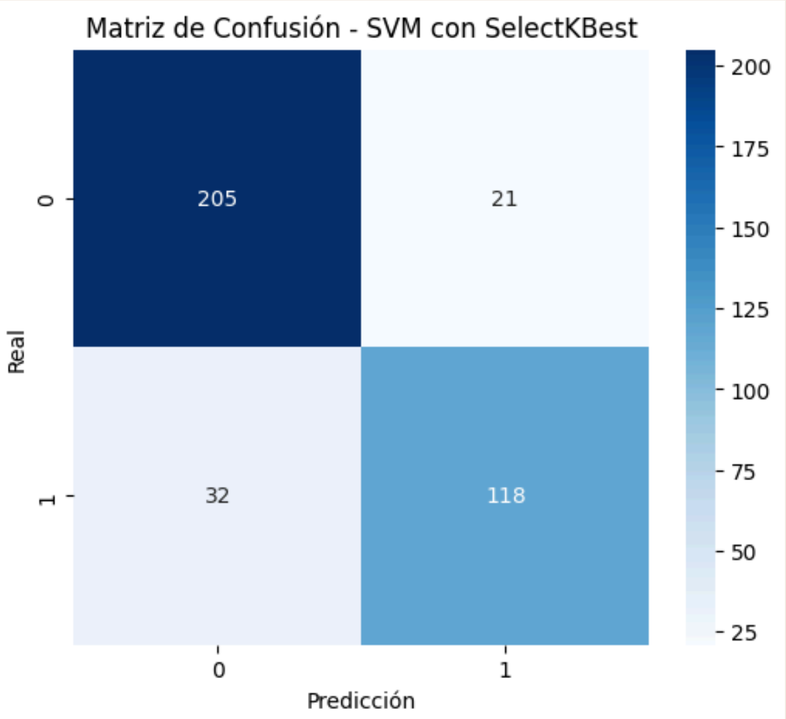


Table 2: Métricas de desempeño del modelo Random Forest con SelectKBest

Clase	Precisión	Sensibilidad	F1-score	Soporte
Clase 0	0.94	0.96	0.95	226
Clase 1	0.94	0.90	0.92	150
Promedio macro	0.94	0.93	0.94	–
Promedio ponderado	0.94	0.94	0.94	–
Precisión global (ACC): 0.94				

Table 3: Métricas de desempeño del modelo Regresion Logistica con SelectKBest

Clase	Precisión	Sensibilidad	F1-score	Soporte
Clase 0	0.86	0.90	0.88	226
Clase 1	0.84	0.79	0.81	150
Promedio macro	0.85	0.84	0.85	–
Promedio ponderado	0.86	0.86	0.86	–
Precisión global (ACC): 0.86				

Table 1: Métricas de desempeño del modelo SVM con SelectKBest

Clase	Precisión	Sensibilidad	F1-score	Soporte
Clase 0	0.86	0.91	0.89	226
Clase 1	0.85	0.79	0.82	150
Promedio macro	0.86	0.85	0.85	–
Promedio ponderado	0.86	0.86	0.86	–
Precisión global (ACC): 0.86				

DISCUSIONES

Mejor modelo: **Random Forest + selección Lasso** (94% accuracy, F1-score 0.93), superando significativamente a modelos lineales ($p<0.01$).
7 predictores clave:

- HbA1c (más relevante).
- Síntomas clásicos (poliuria, polidipsia).
- IMC e hipertensión (factores metabólicos).

Consistente con literatura: métodos ensemble son óptimos para problemas médicos complejos.

Estudio	Método	Accuracy	Características
Presente estudio	**RF + Lasso**	**0.94**	**7**
Alghamdi et al. (2020)	XGBoost	0.91	12
Maniruzzaman et al. (2022)	DL	0.89	15

DISCUSIONES

Ventajas del modelo:

- Mayor precisión con menos variables (gracias a selección rigurosa, optimización de hiperparámetros y datos limpios).

Implicaciones clínicas:

- Tamizaje más simple en atención primaria.
- Reducción de costos (menos pruebas).
- Modelos más interpretables para médicos.

Limitaciones:

- Sesgo de selección (falta diversidad étnica).
- Variables auto-reportadas (ej. dieta).
- Necesidad de validación externa.

Futuras investigaciones:

- Estudios longitudinales.
- Pruebas en entornos reales.
- Análisis de costo-efectividad.

CONCLUSIONES

Este estudio demuestra que:

1. Los modelos de machine learning, particularmente Random Forest con selección Lasso, pueden predecir diabetes con alta precisión (94 %) usando sólo 7 características clínicas clave
2. El conjunto mínimo óptimo incluye: HbA1c, síntomas clásicos (poliuria, polidipsia, pérdida de peso), IMC e hipertensión
3. Este enfoque balancea precisión predictiva con interpretabilidad clínica, facilitando su potencial implementación
4. La validación rigurosa y análisis de errores identificó subgrupos que requerirían ajustes (pacientes medicados con antihipertensivos)

Estos hallazgos apoyan el uso de algoritmos de selección de características combinados con modelos ensemble para desarrollar herramientas diagnósticas parsimoniosas pero precisas en el manejo de diabetes mellitus tipo 2.

MUCHAS GRACIAS

Abril del 2025