Herramientas de Aprendizaje Automático para Tamizaje de Diabetes Tipo 2 en Entornos de Baja Cobertura Médica

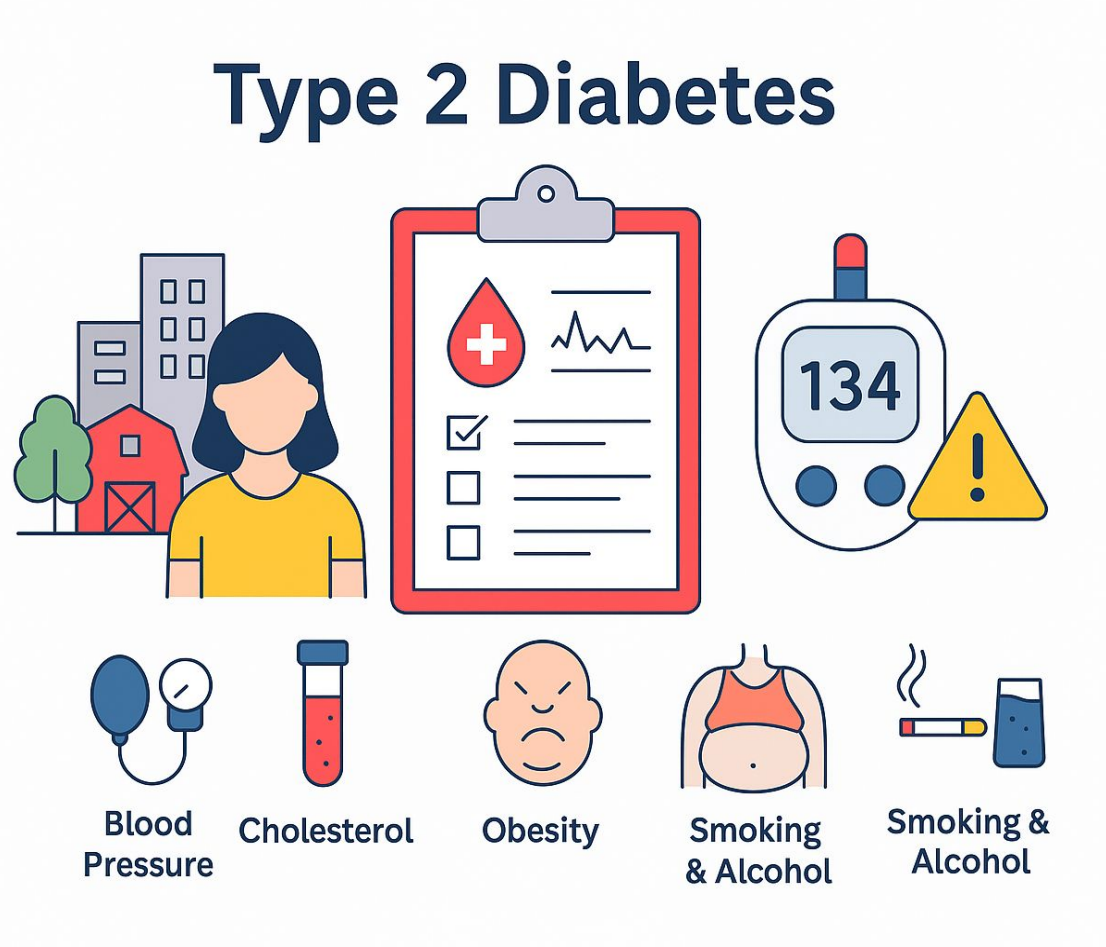
Maestría en Inteligencia Artificial y Ciencia de Datos

Juan Felipe Hernandez
Manuel Enrique Luna
Dallys Nicol Sinisterra Gutierrez

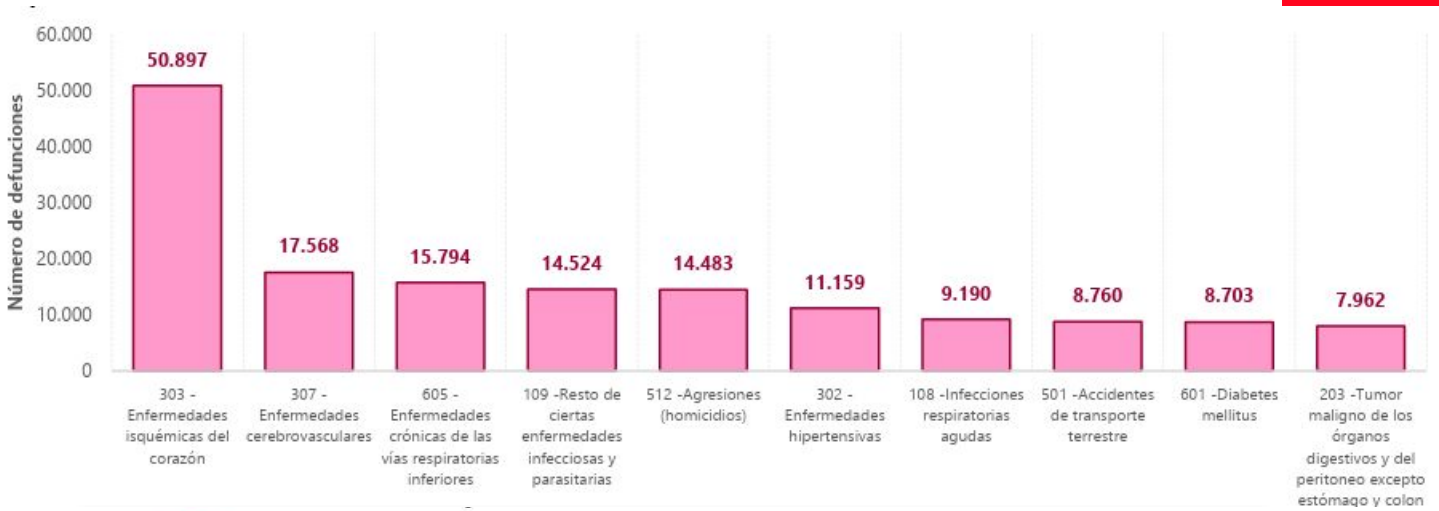
Descripción del Problema

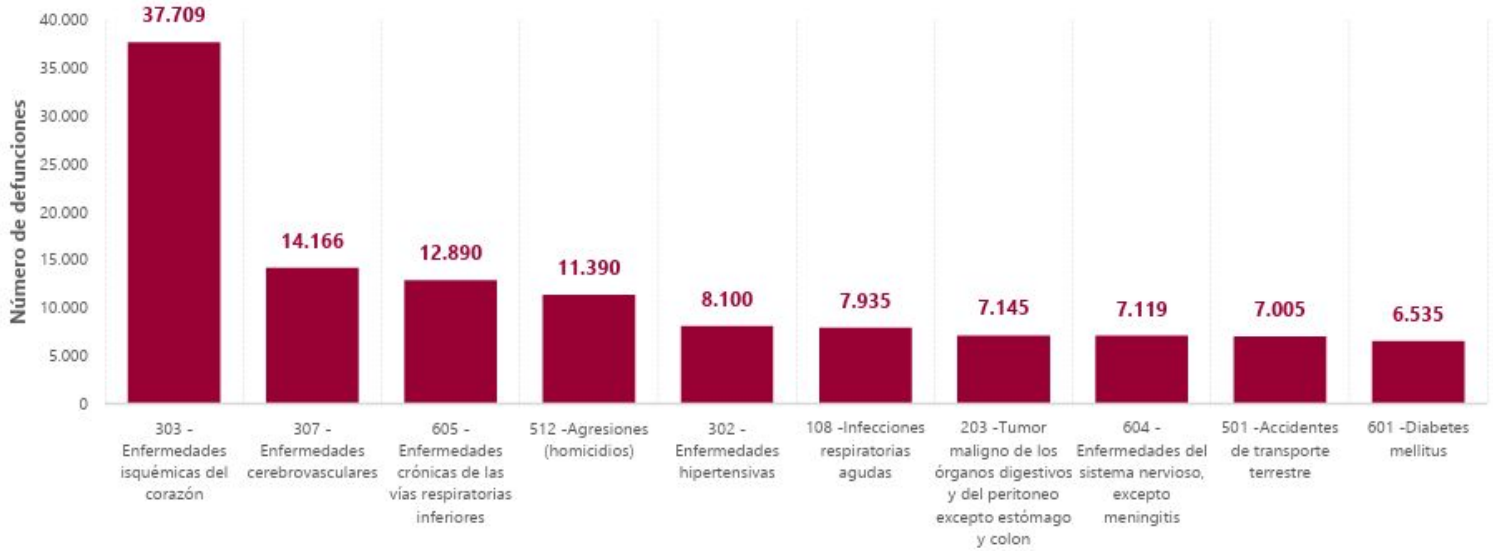
La diabetes tipo 2 es una de las principales enfermedades crónicas no transmisibles que afecta a millones de colombianos, con una tendencia creciente especialmente en zonas rurales y de difícil acceso evidenciando barreras estructurales que limitan el diagnóstico temprano, lo que incrementa complicaciones y eleva los costos para el sistema de salud.

De modo que se propone desarrollar un modelo de clasificación supervisada, basado en machine learning, que permita predecir el riesgo de diabetes tipo 2 usando variables como:
Presión arterial, Colesterol, IMC, Consumo de tabaco y alcohol, Hábitos alimenticios, entre otros.



Impacto del Problema

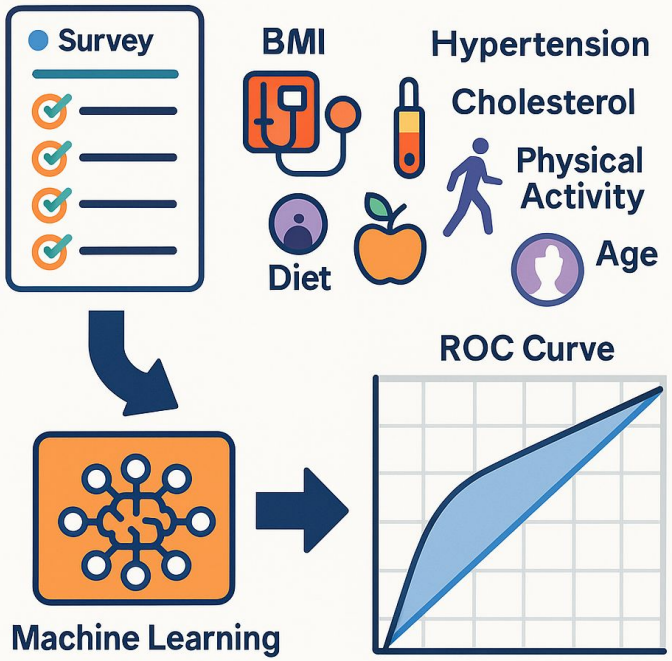




Según las cifras del DANE para 2022, la diabetes mellitus se mantuvo entre las 10 principales causas de muerte en Colombia,

Para el tercer trimestre del 2023 la diabetes mellitus se mantuvo, siendo superior en volumen al del año inmediatamente anterior.

Datos y Metodología



Origen y propósito:

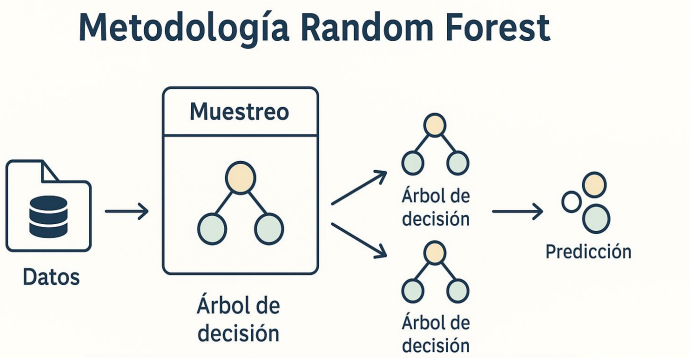
Proviene de la encuesta telefónica anual BRFSS 2015 (Behavioral Risk Factor Surveillance System) de los CDC, destinada a recopilar información sobre factores de riesgo y enfermedades crónicas en adultos en EE.UU.

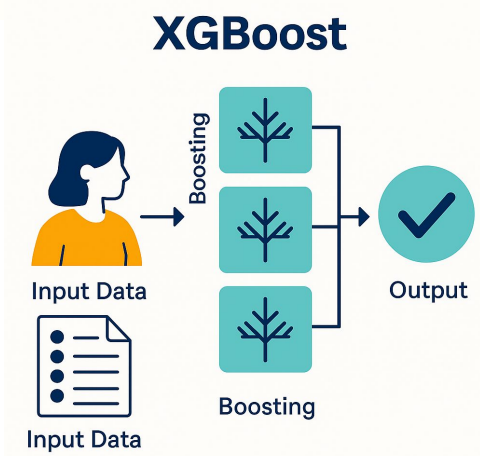
Objetivo del dataset: clasificar a cada encuestado como diabético, pre-diabético o saludable basándose en indicadores de salud auto-reportados.

Tamaño y estructura:

70692 instancias, sin valores faltantes tras limpieza.
21 variables predictoras, incluyendo datos demográficos (edad, sexo), estilo de vida (actividad física, consumo de frutas/verduras), y salud autoinformada (IMC, hipertensión, colesterol alto, salud general, etc.).
Variable objetivo: Diabetes_binary (0 = sano, 1 = pre-diabetes o diabetes)

Aspecto	Estudios previos con datos clínicos	CDC Diabetes Health Indicators (actual)
Variables clave	Glucosa en ayunas, HbA1c, lípidos, fT4, etc.	IMC, hipertensión, colesterol, salud general, hábitos de vida
Profundidad de datos	Datos de laboratorio y EHR (alta precisión)	Auto-reportados (más ruido, pero mayor cobertura y simplicidad)
Tamaño de muestra	Decenas a miles de registros	~~250 000 registros (escala poblacional)
Aplicaciones	Diagnóstico clínico y apoyo a decisiones médicas	Cribado masivo vía encuestas, telemedicina y salud pública
Modelos con mejor AUC	RF, XGBoost con AUC > 0.90 (datos clínicos)	En estudios con datos no invasivos: XGBoost AUC ≈ 0.82; RF ≈ 0.79





Modelos propuestos

Se propone el desarrollo de un modelo de clasificación binaria basado en aprendizaje automático, el cual utilice variables relacionadas con indicadores de salud física,hábitos de vida y condiciones sociodemográficas del paciente, basada en la aplicación de modelos supervisados donde se usará como variable objetivo el campoDiabetes_binary (1 = diabetes, 0 = no diabetes).

Se plantean tres algoritmos principales para la etapa de modelado:

1. ***Regresión logística,*** como modelo base de referencia y por su valor explicativo en contextos de salud, de tal manera que se identifique la razón de prevalencia entre las diferentes categorías.

1. ***Random Forest,*** por su capacidad para manejar relaciones no lineales y su interpretabilidad.

1. ***XGBoost,*** por su alto desempeño en datasets con problemas de clases minoritarias, lo que le atribuye una mayor sensibilidad.

1. **Support Vector Machine,** busca un hiperplano que maximiza el margen entre las dos clases; puede usar “kernels” (lineal, RBF, polinómico) para capturar relaciones no lineales.

Preprocesamiento de los datos

TÉCNICAS UTILIZADAS

- Análisis exploratorio y tratamiento de valores a atípicos
- Escalado de variables
- Selección de características relevantes
- División del conjunto de datos

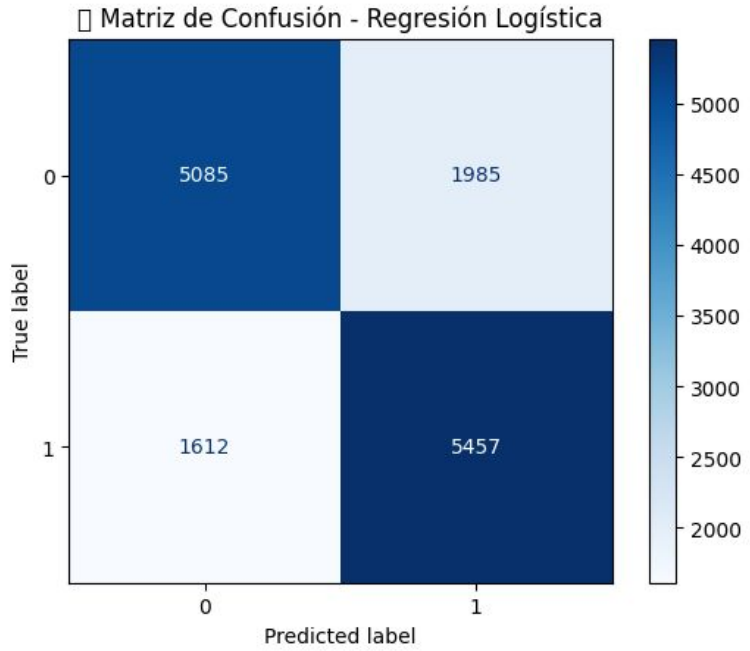
- Se utilizaron diagramas de caja (boxplots), histogramas y funciones de cuantiles para identificar valores extremos en variables continuas como el índice de masa corporal (IMC)

- Para garantizar un entrenamiento balanceado de modelos sensibles a la escala (como SVM y regresión logística), se aplicó el método StandardScaler, que transforma las variables numéricas a una distribución con media cero y desviación estándar uno.

- Se utilizó la técnica SelectKBest con la función de prueba ANOVA para seleccionar las variables más discriminantes con respecto a la variable objetivo

- Una vez preprocesados los datos, se realizó una partición del conjunto en subconjuntos de entrenamiento (80%) y prueba (20%)

Regresión Logística



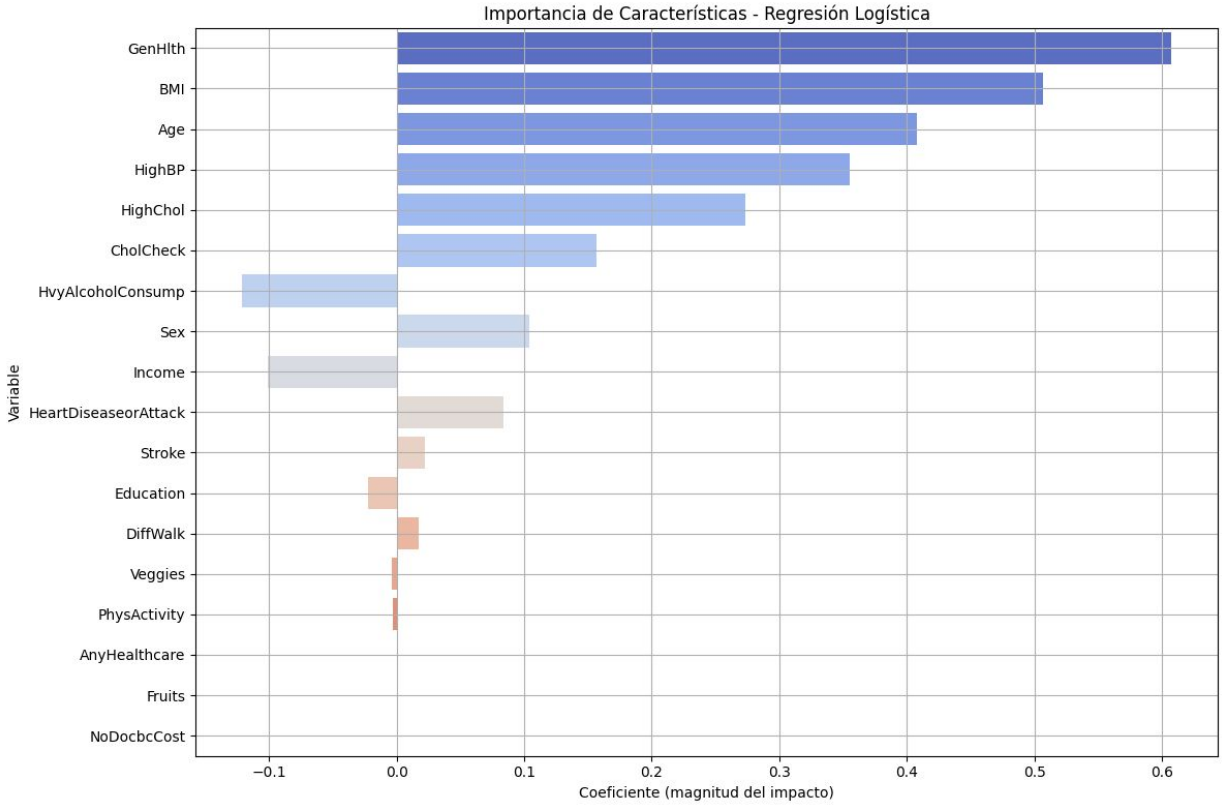
Reporte con variables reducidas:

	precision	recall	f1-score	support
0.0	0.76	0.72	0.74	7070
1.0	0.73	0.77	0.75	7069
accuracy			0.75	14139
macro avg	0.75	0.75	0.75	14139
weighted avg	0.75	0.75	0.75	14139

Accuracy: 0.7455972841077869

El modelo de regresión logística, usando un conjunto reducido de variables, obtuvo una precisión general del 74.55 %, con un desempeño equilibrado en ambas clases. La matriz de confusión muestra que el modelo predijo correctamente 5085 casos negativos (clase 0) y 5457 casos positivos (clase 1), mientras que cometió 1985 falsos positivos y 1612 falsos negativos. El f1-score fue 0.74 para la clase 0 y 0.75 para la clase 1, lo cual indica una buena capacidad del modelo para clasificar ambos grupos de forma relativamente balanceada.

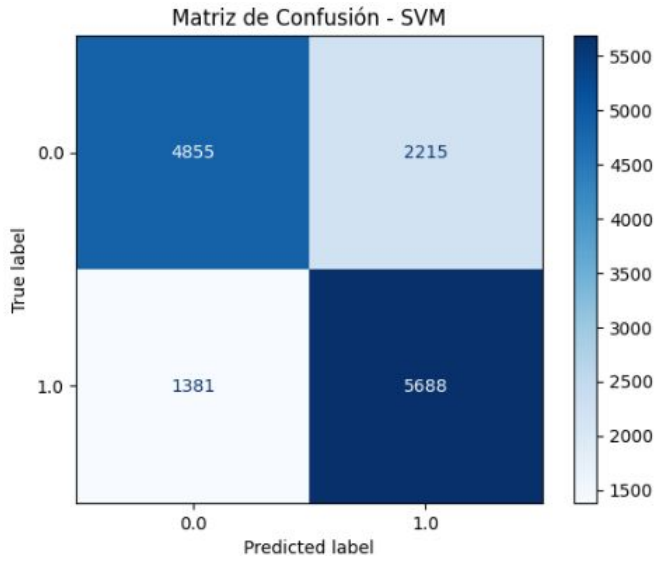
Regresión Logística - Odds Ratios



Las variables con odds ratio mayores a 1 indican que aumentan la probabilidad de que una persona tenga diabetes (clase 1), mientras que las que tienen odds ratio menores a 1 están asociadas con una menor probabilidad. Cuanto más alejado esté el odds ratio de 1, mayor es el impacto (positivo o negativo) de esa variable.

Variable	Coeficiente	Odds Ratio
GenHlth	0.607235	1.835349
BMI	0.506977	1.660264
Age	0.407604	1.503211
HighBP	0.355533	1.426941
HighChol	0.273681	1.314795
CholCheck	0.157003	1.169999
Sex	0.104279	1.109910
HeartDiseaseorAttack	0.083937	1.087560
Stroke	0.022397	1.022650
DiffWalk	0.017126	1.017274
Fruits	0.000000	1.000000
NoDocbcCost	0.000000	1.000000
AnyHealthcare	0.000000	1.000000
PhysActivity	-0.002764	0.997240
Veggies	-0.003991	0.996017
Education	-0.022213	0.978032

Support Vector Machine (SVM)



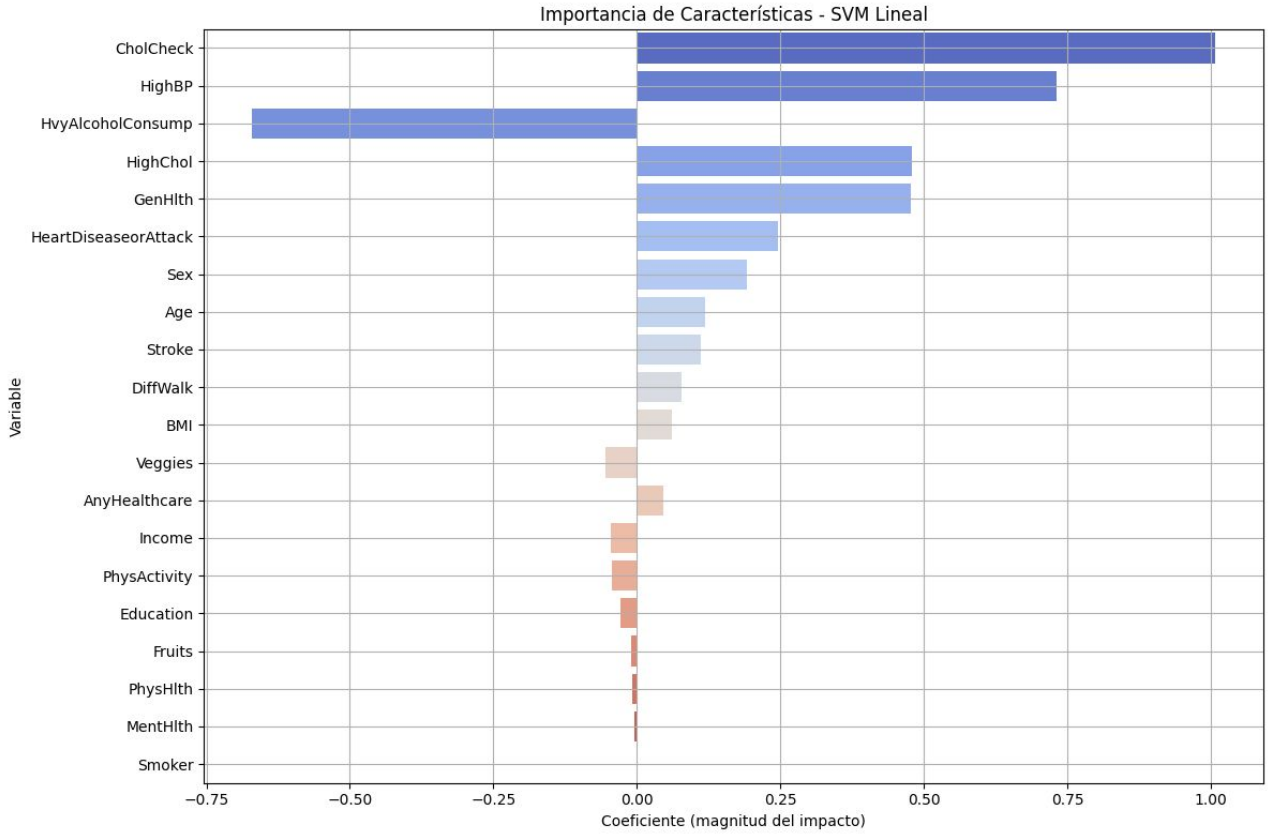
Reporte de Clasificación - SVM sin preprocesamiento:

	precision	recall	f1-score	support
0.0	0.77	0.71	0.74	7070
1.0	0.73	0.79	0.76	7069
accuracy			0.75	14139
macro avg	0.75	0.75	0.75	14139
weighted avg	0.75	0.75	0.75	14139

Accuracy: 0.7470118113020723

El modelo de Máquina de Vectores de Soporte (SVM), aplicado con escalado previo de características continuas, logró un rendimiento equilibrado con una precisión global del 74.7%. Mostró una mayor capacidad para identificar correctamente casos positivos de diabetes (recall de 0.79 en clase 1), lo que lo hace útil en contextos donde es prioritario minimizar falsos negativos. Aunque incurre en algunos falsos positivos, su desempeño general lo posiciona como una opción confiable para tareas de clasificación binaria en este dominio.

Importancia de Características - SVM



La gráfica muestra las 20 características más influyentes en la predicción de diabetes según el modelo de SVM con kernel lineal, basado en los coeficientes del modelo. Las variables con coeficientes positivos (en azul) aumentan la probabilidad de ser clasificado como persona con diabetes, mientras que las de coeficientes negativos (en rojo) disminuyen dicha probabilidad.

Modelo Random Forest con Fine-tuning de hiper-parámetros

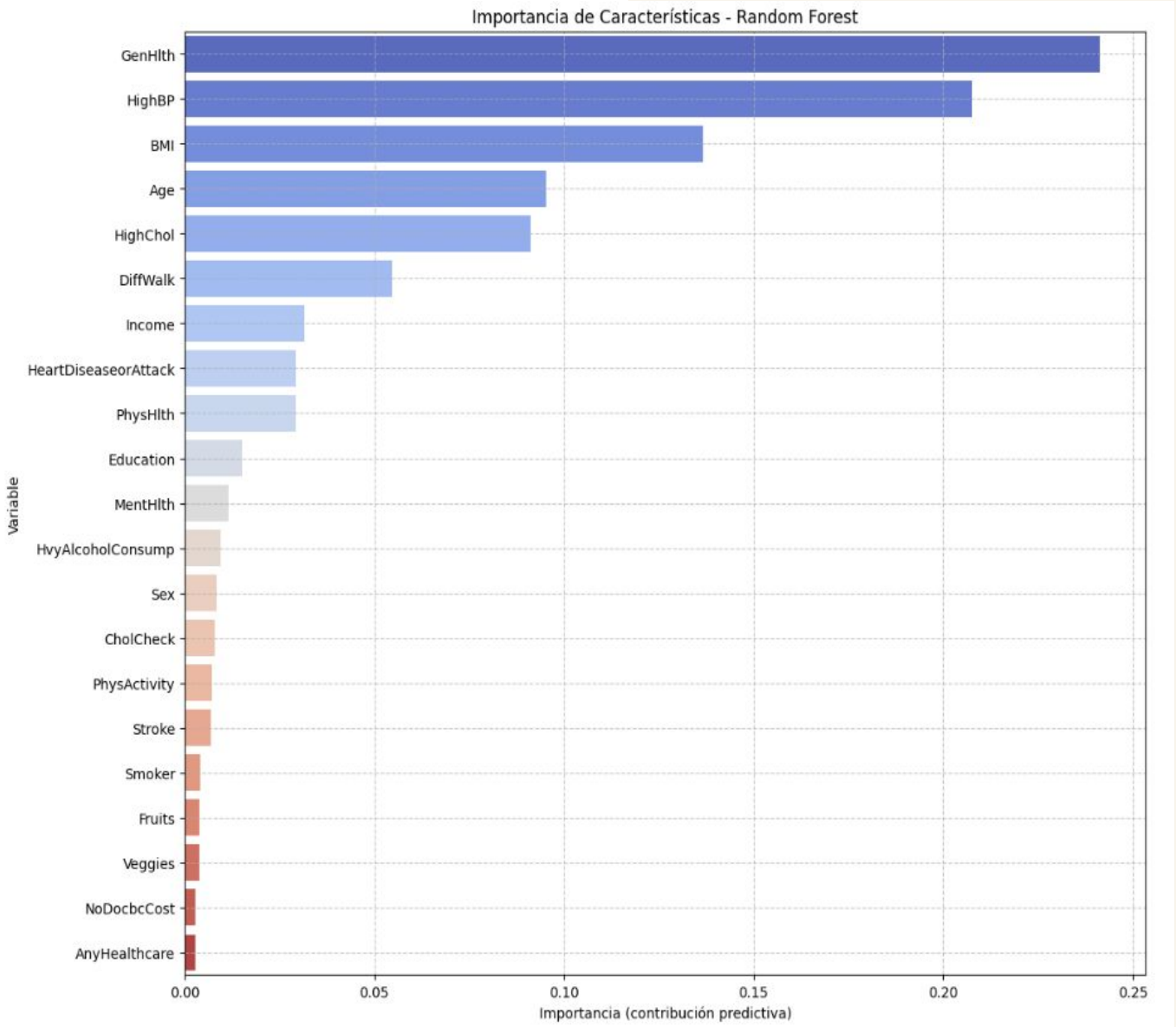
- `max_depth` (Profundidad Máxima): Profundidad máxima de cada árbol.
  - Evita el sobreajuste (overfitting).

- `min_samples_leaf` (Muestras Mínimas por Hoja): Especifica el número mínimo de muestras requeridas para estar en una hoja.
  - Aumentarlo ayuda a que el modelo sea más general y menos sensible al ruido en los datos.

- `n_estimators` (Número de Árboles): Define cuántos árboles componen el Random Forest.
  - Más árboles → menor varianza, pero mayor tiempo de cómputo.

- `class_weight` (Peso de la Clase): Peso asignado a cada clase en la función de pérdida.
  - Mejora el rendimiento en clases minoritarias.

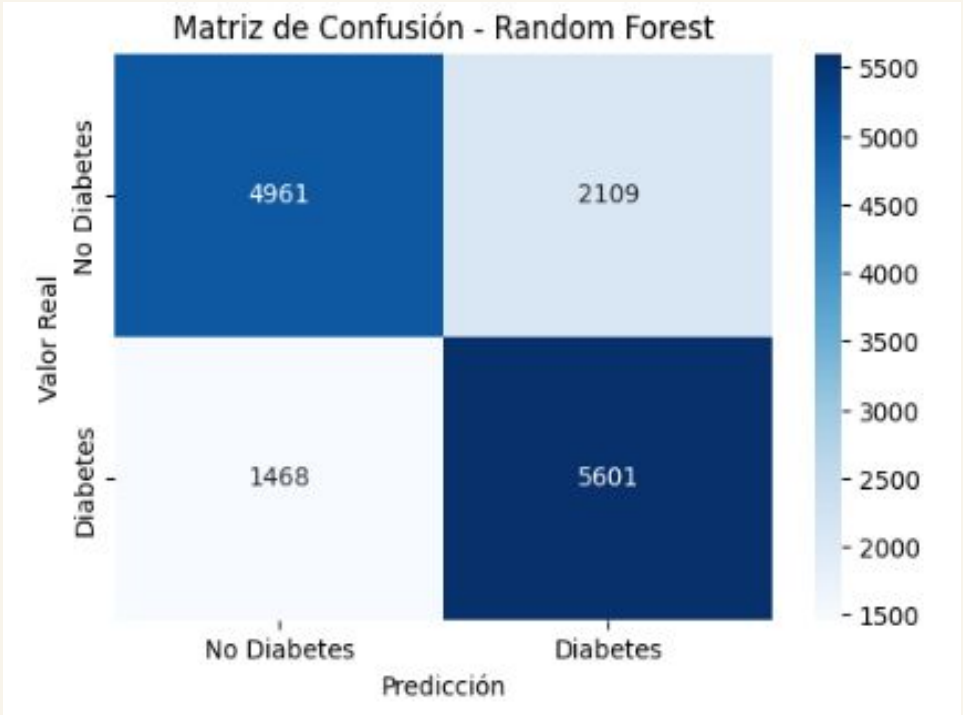
Modelo Random Forest



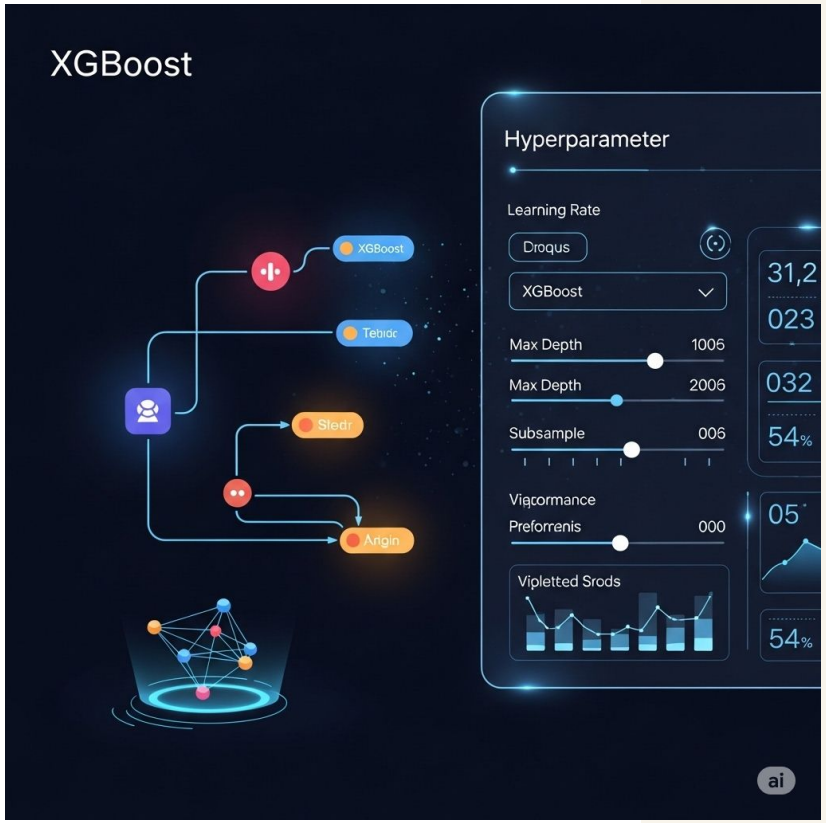
Precisión (Accuracy) del Random Forest : 0.7470

Reporte de Clasificación del Random Forest :

	precision	recall	f1-score	support
0.0	0.77	0.70	0.74	7070
1.0	0.73	0.79	0.76	7069
accuracy			0.75	14139
macro avg	0.75	0.75	0.75	14139
weighted avg	0.75	0.75	0.75	14139



Modelo XGBoost con Fine-tuning de hiper-parámetros



- `learning_rate` (Tasa de Aprendizaje) Controla el tamaño del paso en cada iteración del boosting.
  - Valores bajos (e.g., 0.01–0.1) hacen el modelo más robusto pero requieren más árboles.

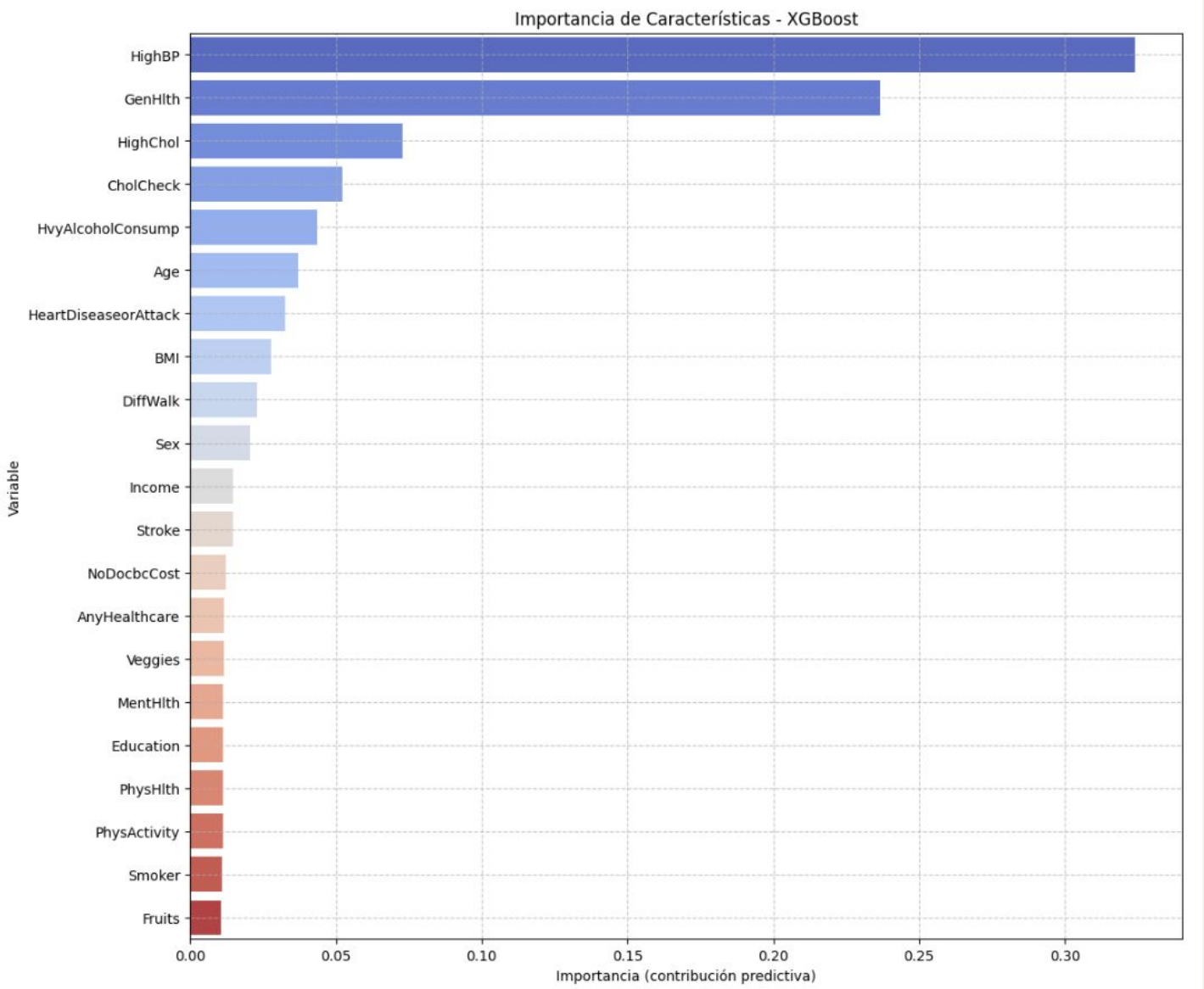
- `max_depth` (Profundidad Máxima): Profundidad máxima permitida para los árboles. Controla la complejidad del modelo.
  - Mayor profundidad → mayor capacidad de ajuste, pero riesgo de overfitting.

- `subsample` (submuestra): Proporción de ejemplos de entrenamiento usados para cada árbol.
  - Introduce aleatoriedad → mejora la generalización.

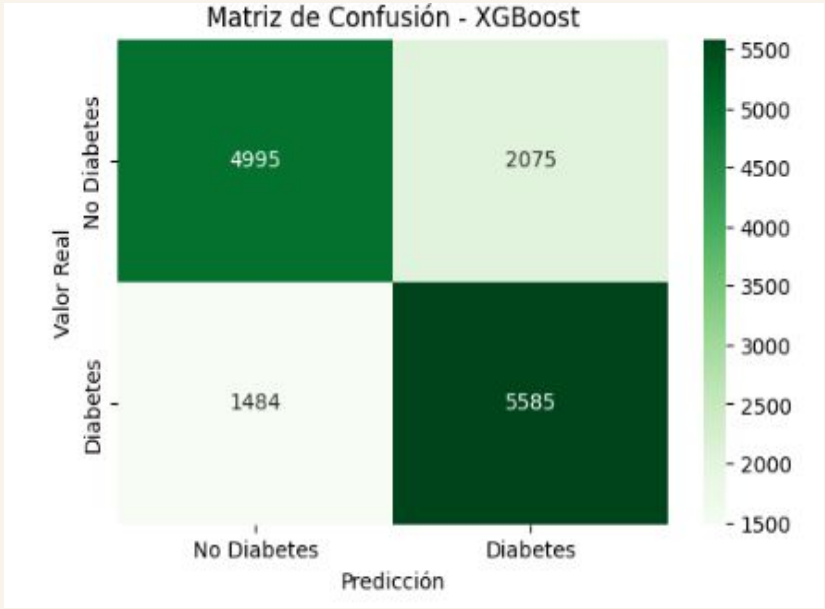
- `colsample_bytree:` Proporción de variables usadas para construir cada árbol.
  - Reduce la correlación entre árboles → mejora la diversidad.

- `scale_pos_weight` (Escala del Peso Positivo): Pondera la clase positiva en problemas desbalanceados.
  - Mejora el recall cuando hay muchas más clases negativas.

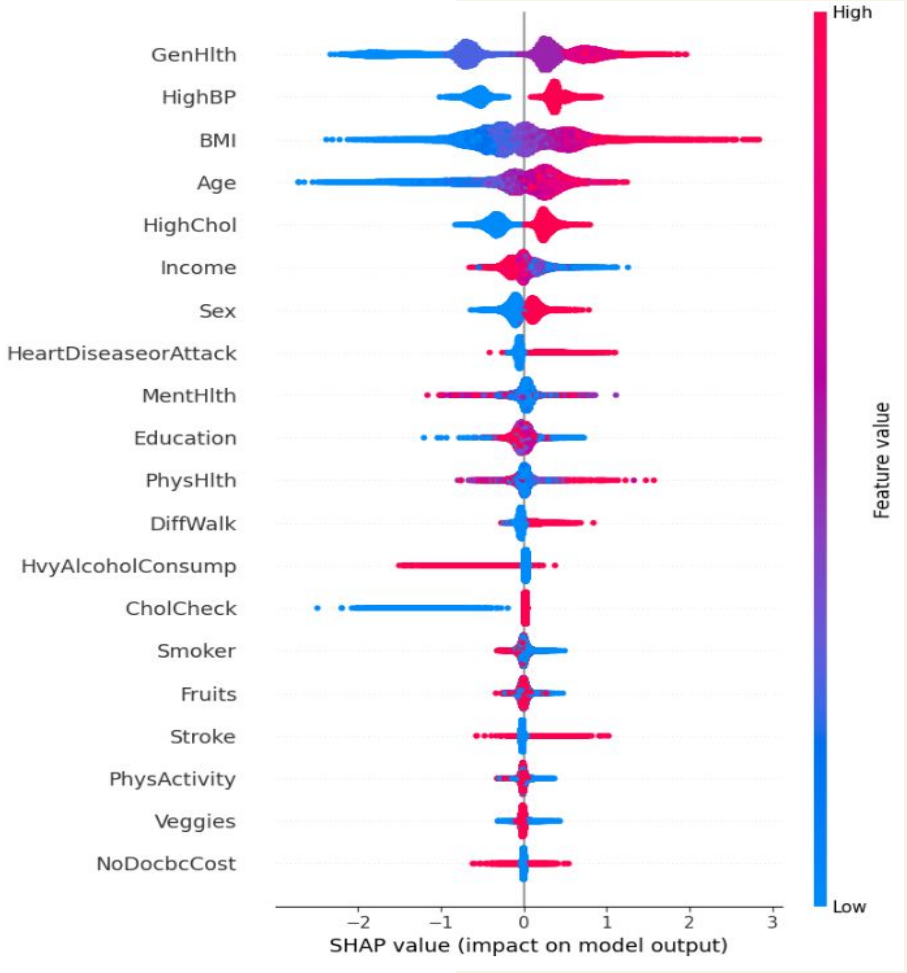
Modelo XGBoost




```
Precisión (Accuracy) del XGBoost: 0.7483

Reporte de Clasificación del XGBoost:
              precision    recall  f1-score   support

         0.0       0.77      0.71      0.74      7070
         1.0       0.73      0.79      0.76      7069

    accuracy                           0.75     14139
   macro avg       0.75      0.75      0.75     14139
weighted avg       0.75      0.75      0.75     14139
```



XGBoost - Valores SHAP (SHapley Additive exPlanations)



- A medida que el valor de **GenHlth** disminuye (de 5=Excelente a 1=Mala salud), su valor SHAP tiende a aumentar significativamente

- Los valores de **GenHlth** de 1 y 2 tienen un fuerte impacto positivo. Inversamente, una mejor salud general (GenHlth 4 o 5) se asocia con valores SHAP negativos.

- **BMI** para un mismo nivel de GenHlth (especialmente en los niveles de peor salud, como GenHlth 1 o 2), los individuos con un BMI más alto (puntos más rojizos/cálidos) tienden a tener valores SHAP aún más elevados, lo que sugiere que la combinación de una mala percepción de la salud general y un alto IMC amplifica la contribución de GenHlth hacia la predicción de diabetes

Model Comparison

	Modelo	Accuracy	Precision Clase 0	Precision Clase 1	Recall Clase 0	Recall Clase 1	F1 Clase 0	F1 Clase 1
0	XGBoost	0.752	0.777	0.731	0.706	0.797	0.740	0.762
1	SVM	0.747	0.769	0.729	0.707	0.788	0.737	0.757
2	Random Forest	0.747	0.772	0.726	0.702	0.792	0.735	0.758
3	Regresión Logística	0.746	0.760	0.733	0.719	0.773	0.739	0.753

El modelo **XGBoost se destaca como el más eficaz**, alcanzando el **Recall más alto con un 79.7%**, lo que significa que es el que mejor logra identificar a los pacientes que realmente padecen la enfermedad, minimizando el riesgo crítico de falsos negativos. Aunque la **Regresión Logística** presenta la **Precisión más alta con un 73.3%**, siendo la más fiable al momento de emitir un diagnóstico positivo, es el F1-Score el que nos da la visión más balanceada. Con un **F1-Score de 0.762**, **XGBoost** demuestra nuevamente su superioridad, logrando el mejor equilibrio general entre la sensibilidad para detectar casos y la fiabilidad de sus predicciones. Por lo tanto, si bien todos los modelos son competitivos, XGBoost se perfila como la opción más robusta y sensible para esta tarea específica.

Conclusiones

XGBoost se posiciona como el modelo con mejor desempeño general para predecir el riesgo de diabetes tipo 2 a partir de variables auto-reportadas, gracias a su equilibrio entre precisión, sensibilidad y F1-score.

Random Forest ofrece resultados comparables con menor complejidad, siendo una opción práctica y eficiente. SVM, aunque preciso para casos negativos, presenta menor sensibilidad, lo que limita su capacidad para detectar casos positivos.

La Regresión Logística, aunque con menor rendimiento, sigue siendo valiosa por su interpretabilidad y adecuada sensibilidad en la clase negativa. Los modelos coinciden en identificar como variables clave la salud general percibida, el IMC, la edad y la presión arterial alta.

Muchas gracias.