Predicción de diabetes a partir de indicadores de salud y estilo de vida

Introducción al Aprendizaje Automático – 1C 2025

Grupo 1

Dasso Julieta Belén
jbdasso@estudiantes.unsam.edu.ar

D'Ercole María Victoria
mvdercole@estudiantes.unsam.edu.ar

Martin Li Gioi Emilio
emartinligioi@gmail.com

Enlace al conjunto de datos original: https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset/data?select=diabetes_binary_health_indicators_BRFSS2015.csv

INTRODUCCIÓN

¿QUÉ ES LA DIABETES?

¡Problemática a nivel mundial!

TIPO I
De origen autoinmune

TIPO II
Asociada a factores como el sobrepeso y el sedentarismo

¿PARA QUÉ INCORPORAR *MACHINE LEARNING*?

¡Para detectar patrones complejos y favorecer la detección temprana de la enfermedad!

OBJETIVO

Desarrollar un modelo predictivo de diagnóstico de diabetes a partir de indicadores de salud y hábitos de vida, utilizando el conjunto de datos *Diabetes Health Indicators.*

¿QUÉ ESPERAMOS ALCANZAR?

Un modelo de *machine learning* que permita clasificar a los pacientes con diabetes a partir de datos de salud y estilo de vida, identificando los factores más relevantes y evaluando distintas alternativas para determinar cuál ofrece el mejor desempeño en la detección temprana de la enfermedad.

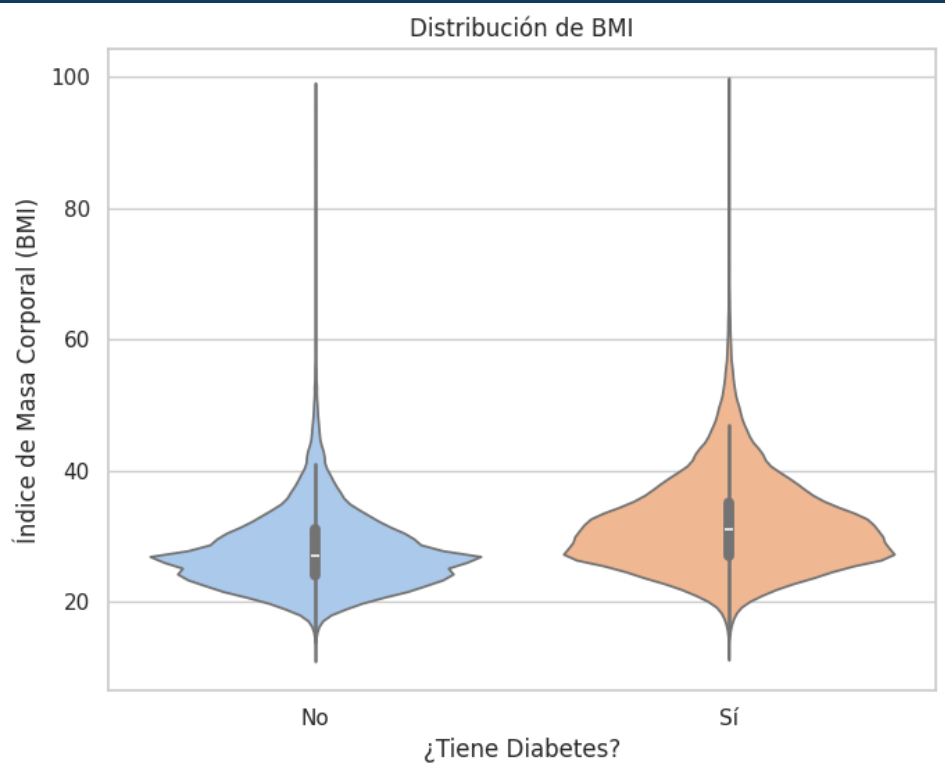
EL DATASET

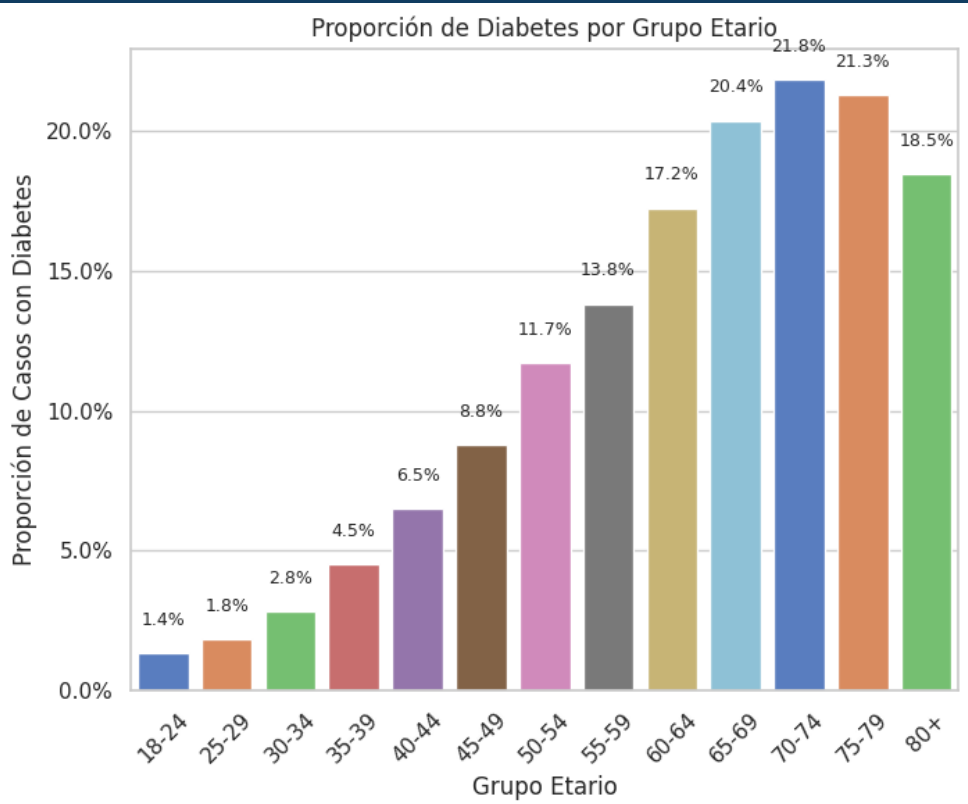
Versión simplificada de una encuesta del programa BRFSS (Sistema de Vigilancia de Factores de Riesgo Conductuales) del CDC (Centros para el Control y la Prevención de Enfermedades) de los Estados Unidos

ASPECTOS GENERALES	VARIABLES DESTACADAS	DESBALANCE DE CLASES
253.680 registros, sin valores faltantes. 4 features continuas, 16 features categóricas, y la variable categórica target (Diabetes_binary).	• GenHlth (Autoevaluación de salud general) • Age (edad) • BMI (índice de masa corporal) • HighBP (hipertensión) • HighChol (colesterol alto) • DiffWalk (dificultad para caminar).	Mayor cantidad de casos de clase negativa (paciente sano, 86.1%) con respecto a la clase positiva (paciente diabético, 13.9%).

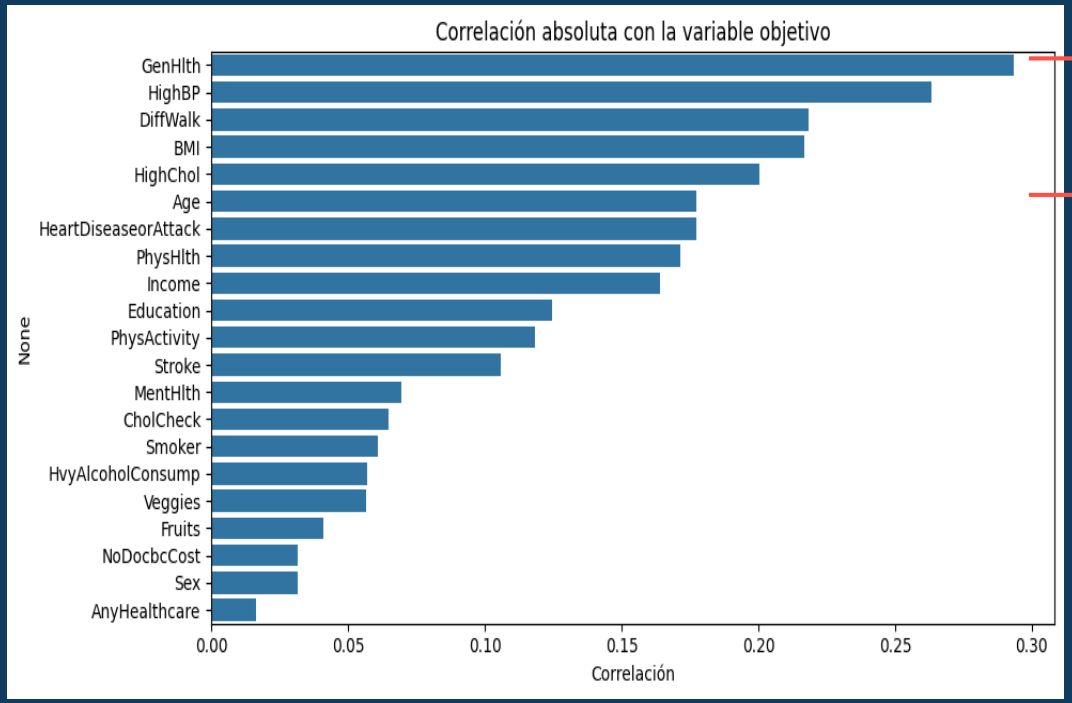
ANÁLISIS EXPLORATORIO

BMI Y AGE





CORRELACIÓN Y SELECCIÓN DE VARIABLES



Adicionalmente a BMI y Age, se incorporan otras 4 variables como posibles factores claves a considerar.

¡ESTO NOS DA EL PIE PARA PENSAR DOS POSIBLES MODELOS DE *MACHINE LEARNING*!

MODELO BENCHMARK

DUMMYCLASSIFIER

El modelo de referencia no realiza aprendizaje real y predice de forma aleatoria según la proporción de clases.

Al evaluarlo, obtuvo solo **14% de F1-score, 14% de precisión** y **14% de recall** para la clase *diabético*, lo que refleja una **capacidad predictiva muy baja**.

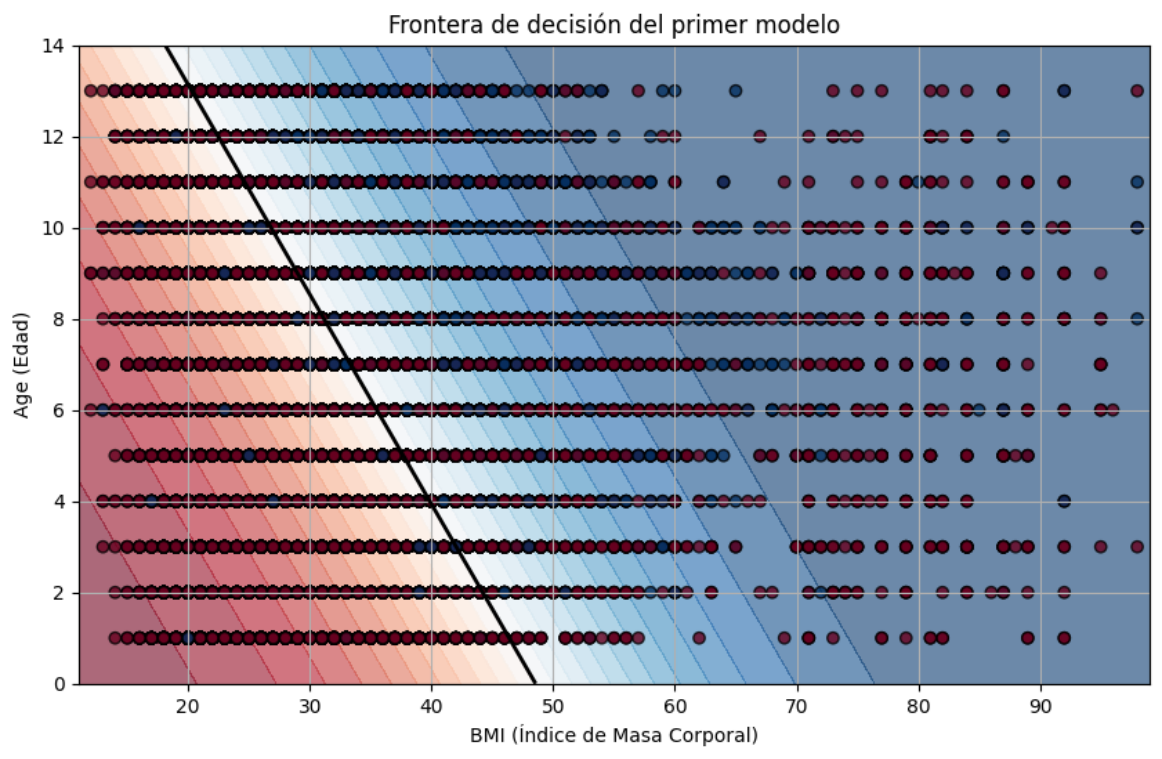
Aun así, cumple su rol como línea base mínima: cualquier modelo más avanzado debe superar claramente estas métricas para ser considerado efectivo

PRIMER MODELO (ML)

Se implementa una regresión logística utilizando dos atributos: BMI y Age, que fueron seleccionados por su relevancia clínica y su fácil disponibilidad.

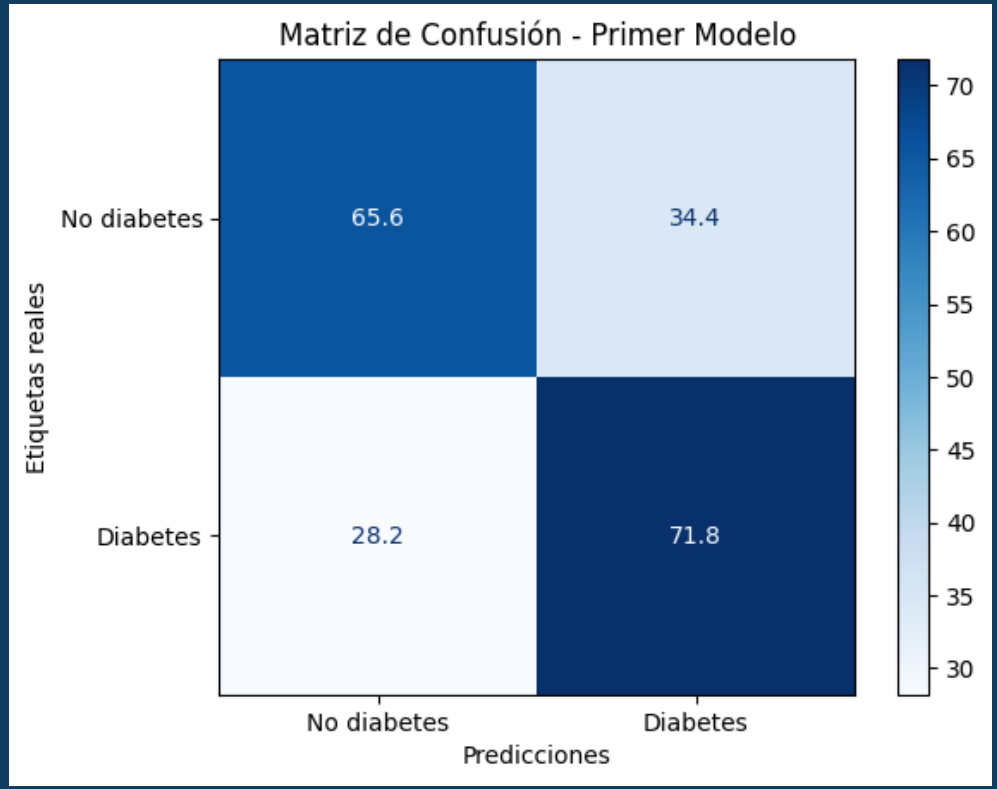
¿CÓMO CLASIFICA EL MODELO VISUALMENTE?

La frontera indica que el modelo predice diabetes incluso con valores moderados de edad y BMI. Esto refleja una estrategia enfocada en **maximizar el recall**, ampliando la detección de posibles casos.



PRIMER MODELO (ML)

El modelo alcanza un **F1-score ponderado de 0.72**, lo que refleja un buen desempeño general.
Este valor equilibra la **alta sensibilidad del modelo (recall del 72%)** con su **baja precisión (25%)**. El modelo **logra detectar la mayoría de los casos reales de diabetes**, aunque a costa de generar varios falsos positivos.



A PARTIR DE ESTOS RESULTADOS...

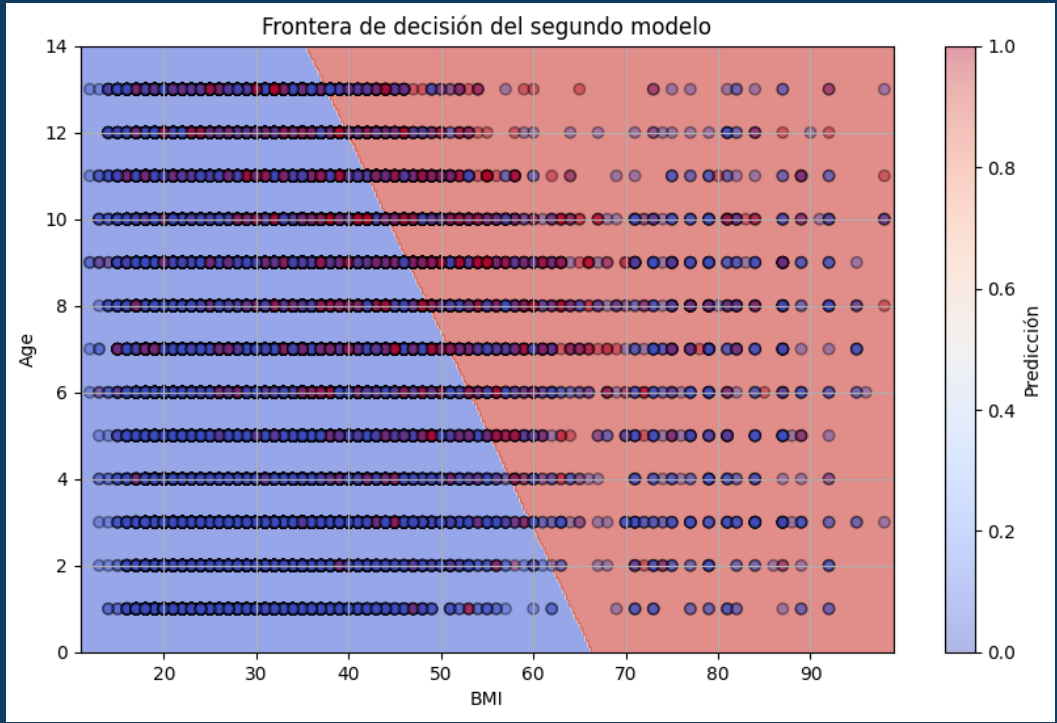
¿Cómo podemos obtener un modelo superador?

SEGUNDO MODELO (ML)

Las variables seleccionadas fueron el BMI, Age, el estado general de salud percibido (GenHlth), la presencia de hipertensión (HighBP), dificultades para caminar (DiffWalk) y colesterol alto (HighChol).

¿CÓMO CLASIFICA EL MODELO VISUALMENTE?

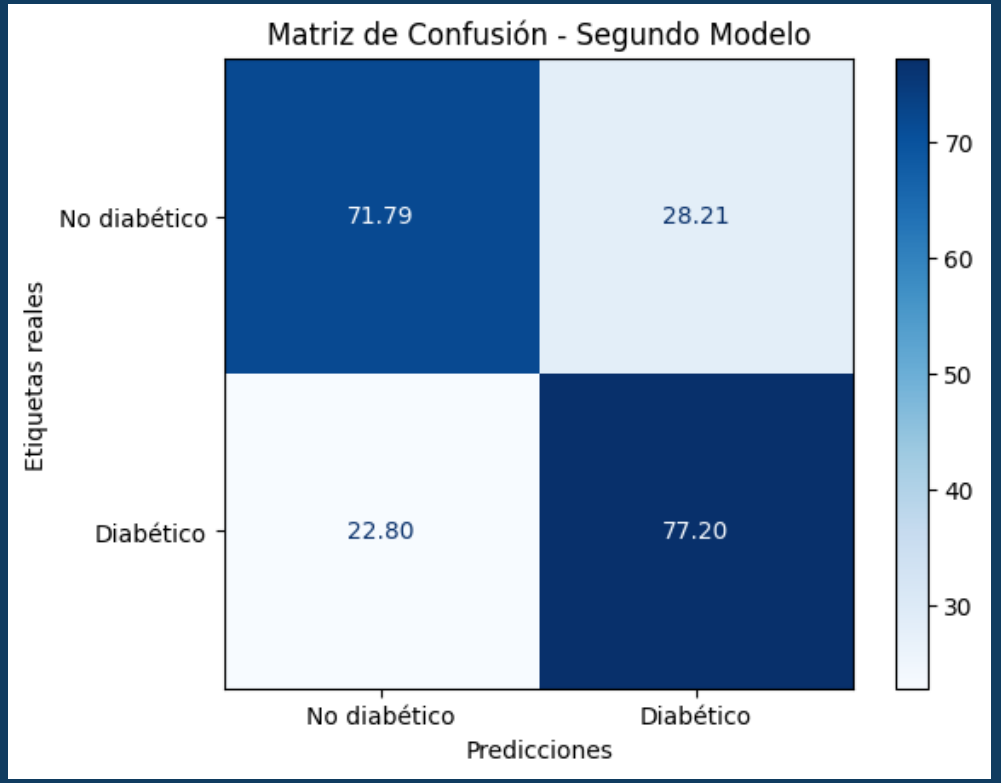
A medida que aumentan el BMI y la edad, el modelo tiende a clasificar con mayor probabilidad como positivos (diabéticos), lo cual es consistente con el conocimiento médico: el riesgo de diabetes suele incrementarse con la edad y el sobrepeso. La decisión del modelo cambia progresivamente en función del perfil del paciente.



SEGUNDO MODELO (ML)

El modelo alcanza un **F1-score ponderado de 0.77** que representa un buen desempeño general superando, en principio, al modelo anterior. Adicionalmente se obtiene un recall para la clase positiva del 77%, acompañado de una precisión del 30%. Esto indica que **el modelo permite identificar la mayoría de los casos reales de pacientes diabéticos**, aunque nuevamente a costa de generar falsos positivos.

En términos generales las métricas obtenidas evidencian un mejor desempeño que su modelo predecesor.



¿CÓMO PODEMOS CONTINUAR TRABAJANDO A PARTIR DE ESTE MODELO?

VALIDACIÓN CRUZADA

Se aplicó **validación cruzada estratificada de 5 folds** sobre el segundo modelo para estimar su capacidad de generalización sin depender de una única partición del dataset.

Las métricas promedio obtenidas fueron:

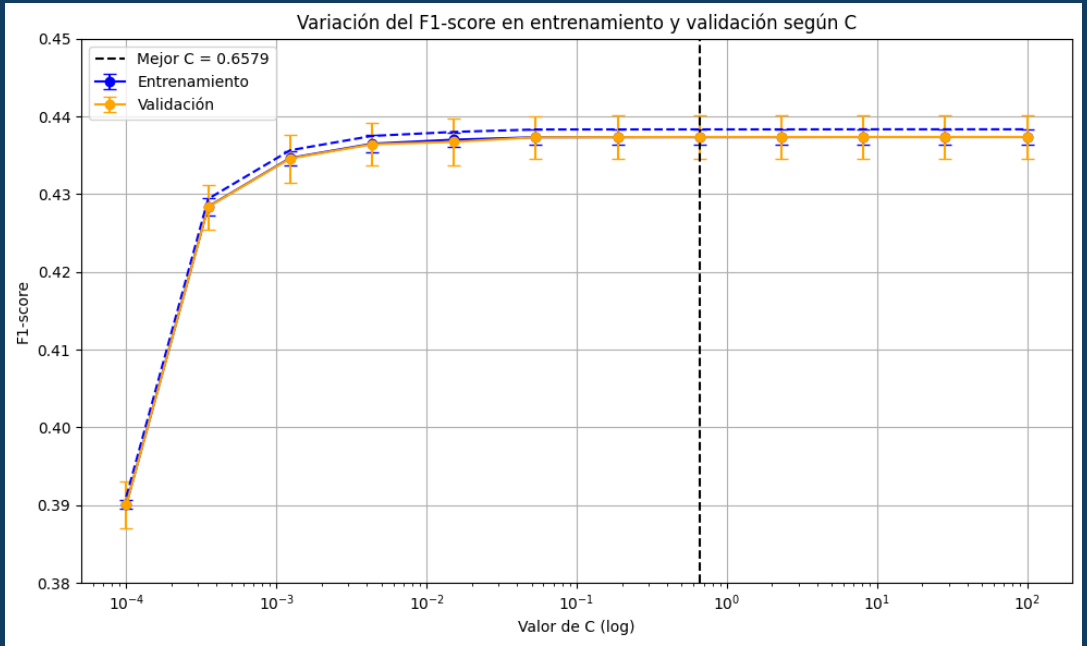
- **Precisión**: 30.5% (± 0.23%)
- **Recall**: 76.2% (± 1.0%)
- **F1-score (clase positiva)**: 43.6% (± 0.38%)

- ❖ Estas métricas coinciden con las obtenidas en el test set, lo que confirma que el modelo es estable.
- ❖ El bajo desvío estándar confirma que el modelo es consistente y confiable en distintas particiones del dataset.

Análisis de sesgo-varianza

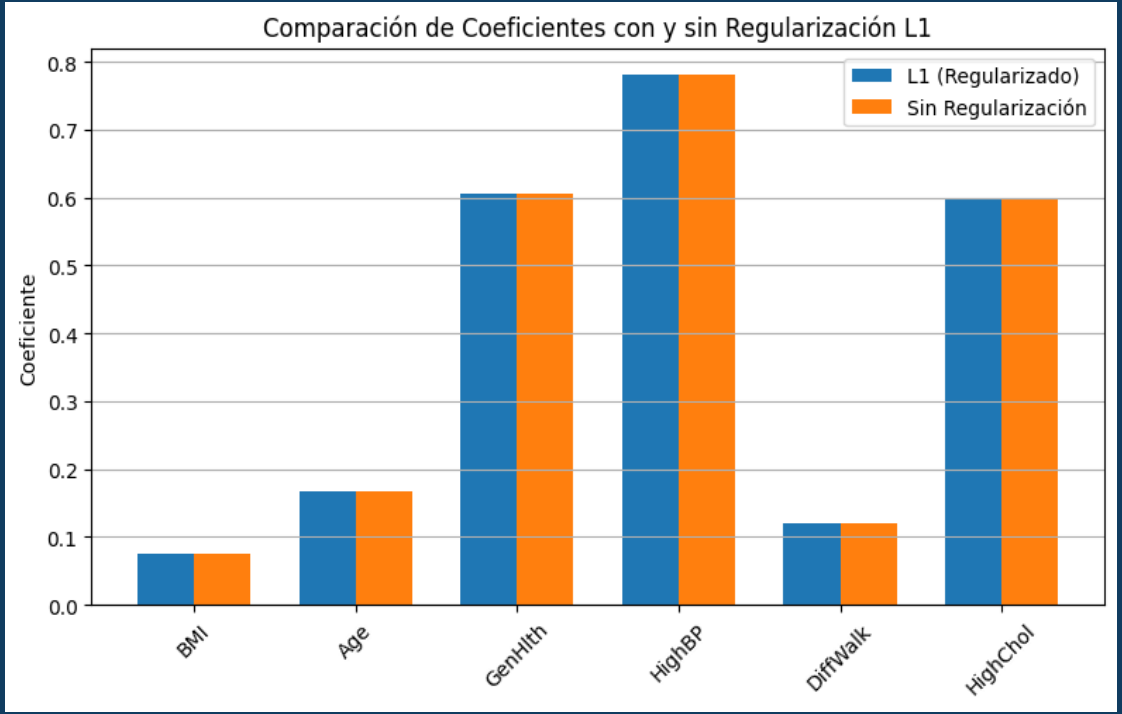
con GridSearchCV y regularización L1

- Para valores bajos de C, el modelo sufre subajuste (F1 ≈ 0.39).
- A partir de C ≈ 0.05, el F1-score se estabiliza cerca de 0.437.
- Curvas de entrenamiento y validación muy similares → baja varianza.
- Mejor valor: C = 0.6579, pero valores mayores (como C = 10) dan el mismo rendimiento.



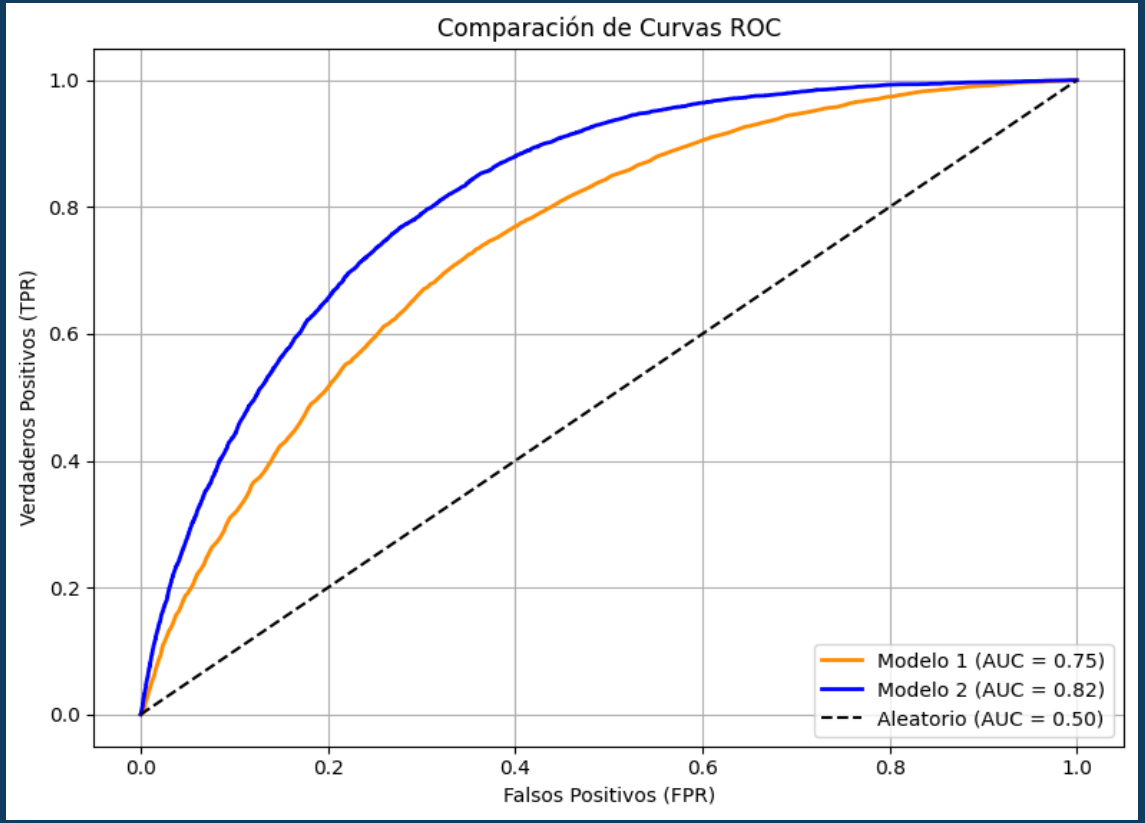
COMPARACIÓN DE COEFICIENTES
CON Y SIN REGULARIZACIÓN

- Los coeficientes del modelo con L1 y sin regularización son prácticamente idénticos.
- Ninguna variable fue anulada por la penalización L1.
- Esto indica que las 6 variables seleccionadas son todas relevantes.
- La regularización no modifica el modelo, pero aporta estabilidad.



COMPARACIÓN DE MODELOS

¿QUÉ PRIORIZAMOS?



El modelo 2 mostró el mejor desempeño entre los modelos evaluados, superando al benchmark y al Modelo 1 en todas las métricas clave. Alcanzó un **AUC de 0.82**, lo que refleja una mayor capacidad para discriminar entre clases.

Si bien es el más adecuado para una implementación práctica, aún existen oportunidades para seguir mejorando su rendimiento.

CONCLUSIONES

Este trabajo exploró la predicción de diabetes a partir de datos de salud y estilo de vida mediante técnicas de *machine learning*. Se comenzó con un modelo de referencia (*dummy*) y se avanzó desde una regresión logística simple hasta un modelo más complejo con seis variables relevantes. Se aplicaron buenas prácticas como el escalado y el balanceo de clases, y se evaluó el rendimiento utilizando métricas como precisión, recall y AUC. Los resultados mostraron que al incorporar más información, complejizando el modelo, las métricas mejoraron considerablemente.

EN CUANTO AL CRITERIO UTILIZADO...

Destacamos la importancia de obtener un aumento de TPR aunque dicho avance pudiera llevar a la obtención de más FP, dado que en el marco de la salud resulta fundamental evitar descartar casos de posible enfermedad.

TRABAJO A FUTURO

¿QUÉ MÁS PODRÍAMOS HACER?

Para continuar desarrollando este modelo, se propone entrenar un modelo de regresión logística utilizando la totalidad de las variables disponibles del dataset. Esto permitiría evaluar si la inclusión de atributos adicionales mejora la capacidad predictiva del modelo y, al mismo tiempo, observar con mayor claridad el efecto de la regularización L1 en contextos con mayor dimensionalidad y potencial redundancia de variables.

También sería interesante explorar modelos más complejos como Random Forest o Gradient Boosting que podrían capturar relaciones no lineales entre los atributos y ofrecer mejoras en métricas como el F1-score para la clase minoritaria (diabéticos).

¡MUCHAS GRACIAS!

¿CONSULTAS?