# Clasificación Supervisada para Predicción de Diabetes

## Introducción

La diabetes representa uno de los desafíos de salud pública más significativos del siglo XXI, afectando a más de 400 millones de personas a nivel mundial según la Organización Mundial de la Salud (OMS). Esta enfermedad crónica, caracterizada por niveles elevados de glucosa en sangre debido a deficiencias en la producción o acción de la insulina, puede llevar a complicaciones graves como enfermedades cardiovasculares, neuropatía, nefropatía y retinopatía si no se detecta y maneja oportunamente.

La detección temprana se convierte en un factor crítico para mejorar los resultados de salud y reducir tanto los costos médicos como el impacto en la calidad de vida de los pacientes. Tradicionalmente, el diagnóstico se basa en pruebas de laboratorio específicas y evaluación clínica de síntomas, un proceso que puede resultar costoso y no siempre accesible para todas las poblaciones, especialmente en áreas con recursos limitados.

La inteligencia artificial, y específicamente las técnicas de aprendizaje automático, ofrecen una oportunidad única para revolucionar este proceso diagnóstico. Estas metodologías pueden identificar patrones complejos en datos clínicos que podrían no ser evidentes para el análisis humano tradicional, permitiendo desarrollar herramientas de screening más eficientes y precisas.

Los árboles de decisión destacan por su interpretabilidad y capacidad para manejar tanto variables numéricas como categóricas, características especialmente valiosas en el contexto médico, donde la comprensión del proceso de toma de decisiones del modelo es fundamental para ganar la confianza de los profesionales de la salud.

El presente trabajo explora cómo las técnicas de clasificación supervisada pueden contribuir al desarrollo de herramientas de apoyo diagnóstico para la diabetes, aprovechando la digitalización actual de la información de salud para transformar la práctica clínica y optimizar la identificación temprana de pacientes en riesgo.

## Objetivos

### Objetivo General

Aplicar técnicas de clasificación supervisada para desarrollar un modelo predictivo que permita identificar la presencia de diabetes en pacientes basándose en indicadores clínicos y demográficos específicos.

### Objetivos Específicos

* **Técnico:** Implementar un algoritmo de árboles de decisión con criterio de éxito de precisión mínima del 70% en el conjunto de datos de prueba.
* **Metodológico:** Realizar preprocesamiento riguroso del dataset, abordando el tratamiento de valores faltantes y valores atípicos.
* **Comparativo:** Evaluar el rendimiento del árbol de decisión contra Random Forest, Regresión Logística y Naive Bayes utilizando métricas estándar.
* **Interpretativo:** Analizar la importancia relativa de las variables predictoras, identificando qué características clínicas contribuyen más significativamente a la predicción.
* **Aplicado:** Demostrar la viabilidad práctica mediante predicciones sobre casos individuales.

## Descripción de Datos Utilizados

Se utilizó el conjunto de datos "Pima Indian Diabetes Database" del Instituto Nacional de Diabetes y Enfermedades Digestivas y Renales de Estados Unidos, enfocado en mujeres de herencia indígena Pima de al menos 21 años de edad.

### Características Estructurales

El dataset comprende 768 registros individuales con 9 variables: 8 características predictoras y 1 variable objetivo para clasificación binaria.

### Variables Predictoras

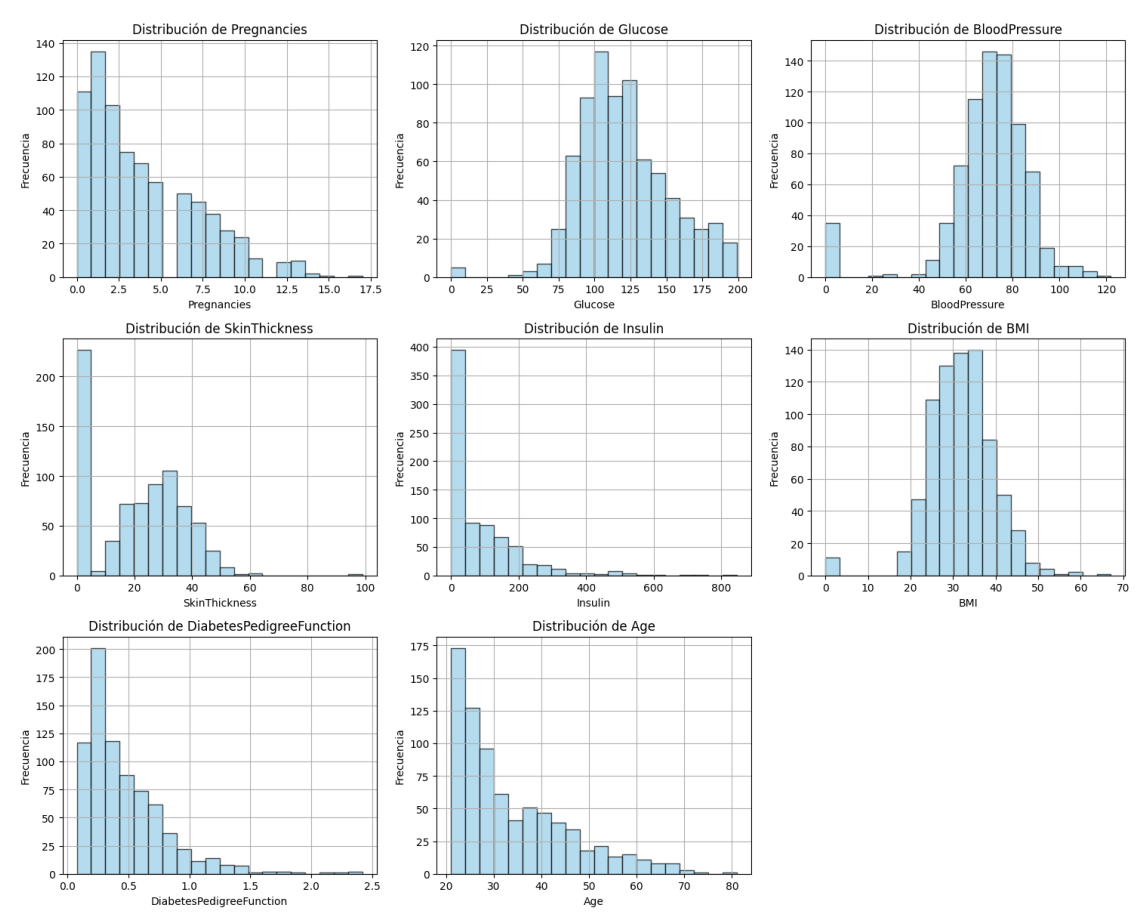
Las variables se organizan en cuatro categorías complementarias del riesgo diabético:

**Variables demográficas y reproductivas:** edad y número histórico de embarazos, siendo este último relevante dado que la diabetes gestacional constituye un factor de riesgo para diabetes tipo 2.

**Indicadores metabólicos directos:** concentración de glucosa plasmática post-carga (mg/dL) y niveles séricos de insulina (mu U/ml), capturando aspectos centrales de la fisiopatología diabética.

**Mediciones antropométricas y fisiológicas:** índice de masa corporal (BMI), espesor del pliegue cutáneo del tríceps y presión arterial diastólica, relacionados con resistencia insulínica y síndrome metabólico.

**Variable genética:** "Diabetes Pedigree Function", que cuantifica la probabilidad de diabetes basándose en historia familiar, incorporando predisposición genética.



### Variable Objetivo y Distribución

La variable dependiente es binaria (0=sin diabetes, 1=con diabetes), con distribución de 500 casos negativos (65.1%) y 268 casos positivos (34.9%). Esta prevalencia elevada es consistente con estudios epidemiológicos en la población indígena Pima.

### Limitaciones de los Datos

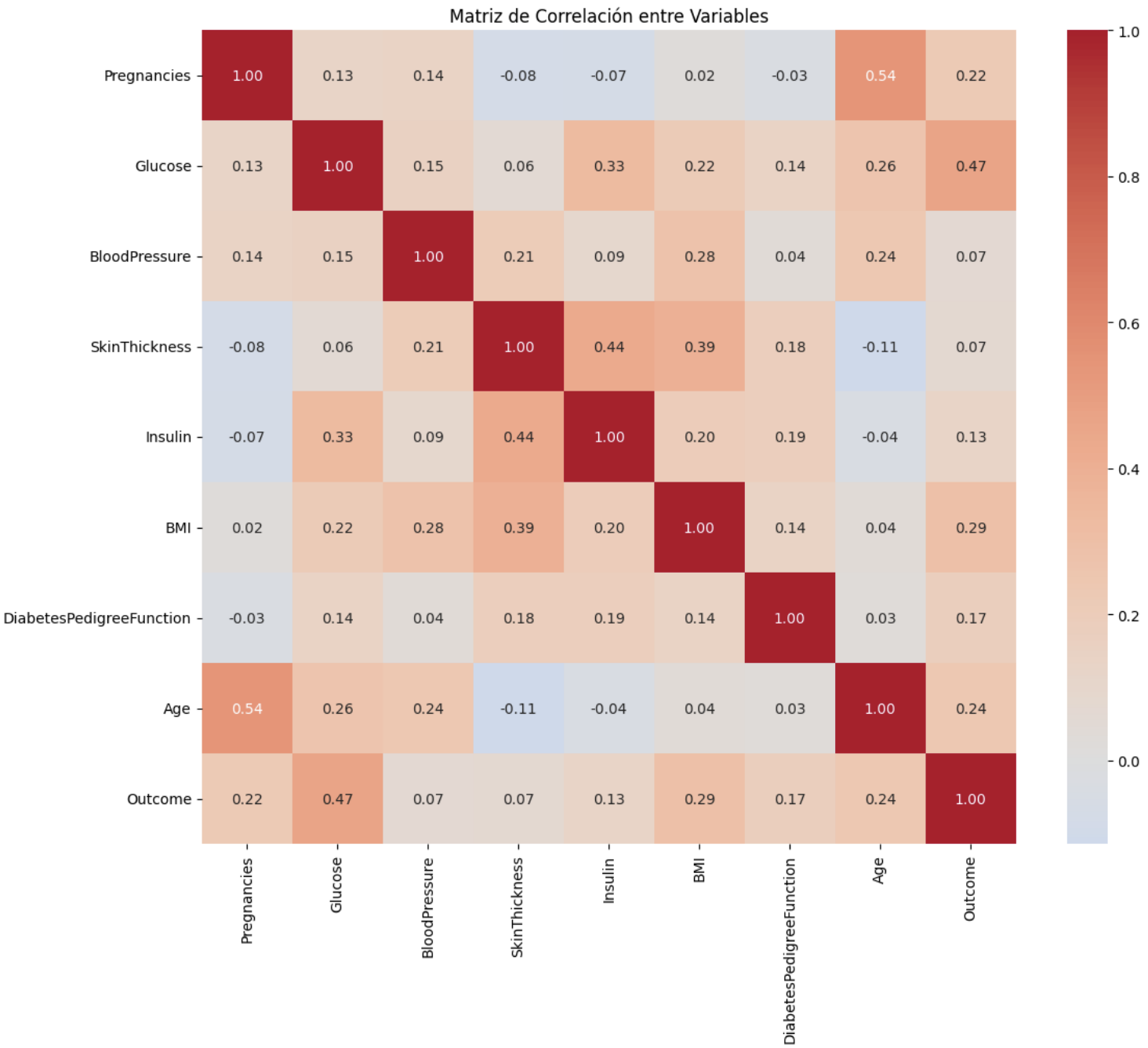
Se identificaron valores cero en variables donde no deberían existir: glucosa (5 casos), presión arterial (35 casos), espesor del pliegue cutáneo (227 casos), insulina (374 casos) y BMI (11 casos), probablemente representando datos faltantes codificados incorrectamente.

## Metodología

Este estudio implementó un enfoque de aprendizaje automático supervisado estructurado en cinco fases principales:

### Análisis Exploratorio

Se calcularon estadísticas descriptivas y se generaron visualizaciones clave, incluyendo matriz de correlación, mapas de calor y matrices de dispersión para identificar patrones y relaciones entre variables.



### Preprocesamiento

Los valores cero problemáticos se transformaron a datos faltantes e imputaron usando la media aritmética de cada variable, estrategia seleccionada por su interpretabilidad y simplicidad.

### División y Validación

División aleatoria estratificada 70-30 para entrenamiento y prueba, manteniendo la distribución de clases en ambas particiones.

### Configuración del Modelo

El árbol de decisión se configuró con parámetros de regularización: profundidad máxima de cinco niveles, mínimo de 20 muestras por división y 10 por hoja, balanceando capacidad expresiva y generalización.

### Evaluación

Se compararon cuatro algoritmos usando métricas múltiples (precisión, sensibilidad, especificidad, F1-score) y matrices de confusión, complementado con análisis de importancia de variables.

## Resultados Obtenidos y Visualizaciones Clave

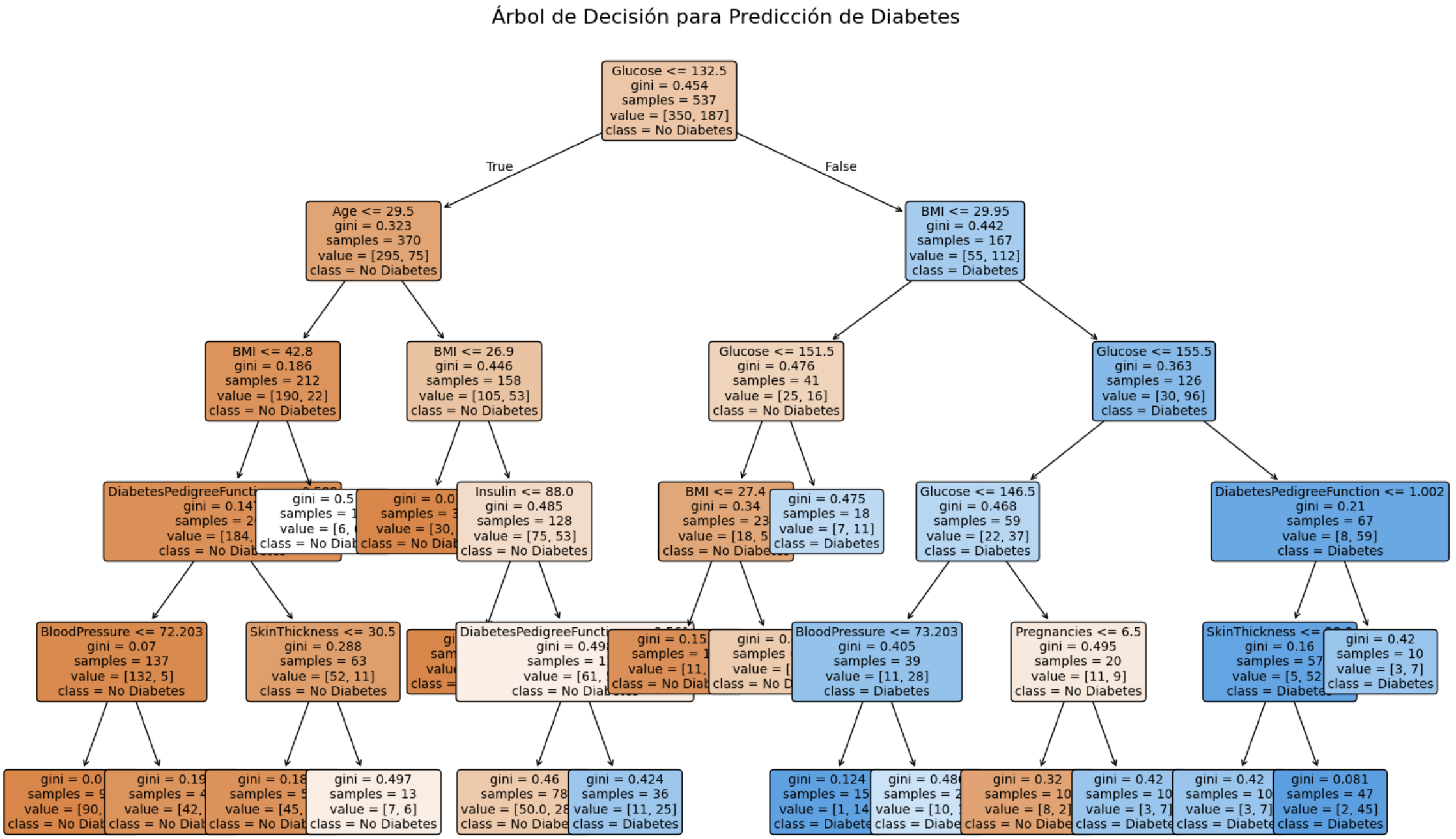
### Rendimiento del Modelo

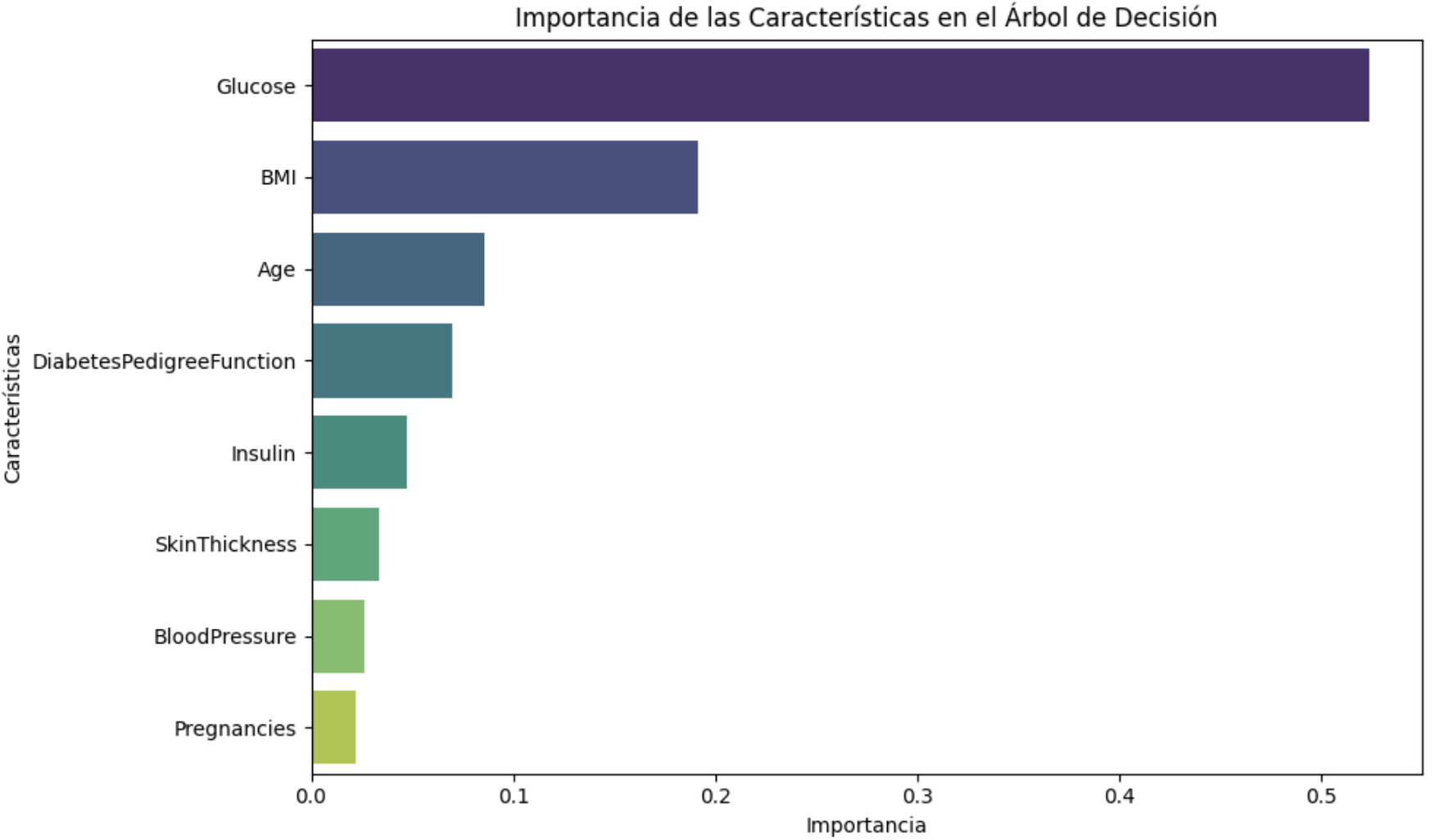
El árbol de decisión alcanzó 76.19% de precisión en prueba, superando el criterio establecido. Mostró mayor especificidad (87%) que sensibilidad (57%), comportamiento conservador con 35 falsos negativos como principal desafío.

### 

### Importancia de Variables

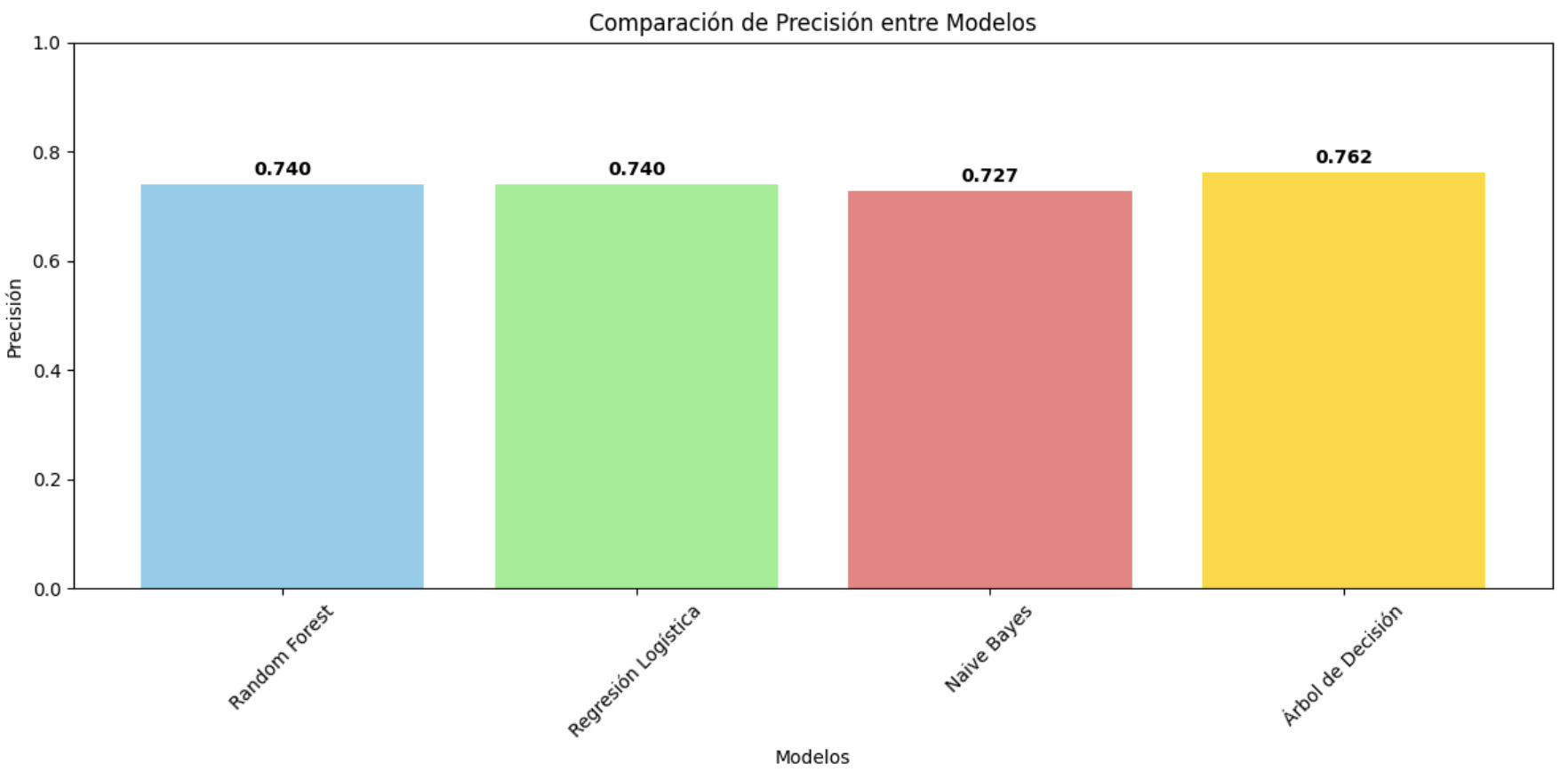
La glucosa contribuyó con 52.36% de la capacidad predictiva, seguida por BMI (19.18%) y edad (8.56%). La función de pedigree diabético aportó 7.01%, confirmando hallazgos clínicamente coherentes con indicadores fácilmente obtenibles en atención primaria.

****



### Comparación de Algoritmos

El árbol de decisión superó a Random Forest y regresión logística (74.03%) y Naive Bayes (72.73%) por su capacidad de capturar interacciones no lineales manteniendo interpretabilidad.



### Visualizaciones Clave

La matriz de correlación confirmó que glucosa presenta la correlación más fuerte (0.467), seguida por BMI (0.293) y edad (0.238). La visualización del árbol mostró que la primera división utiliza un umbral de glucosa de 127.5 mg/dL, convergiendo con el criterio diagnóstico clínico de 126 mg/dL.

### Significancia Clínica

Los resultados demuestran que herramientas de apoyo diagnóstico con precisión clínicamente relevante pueden desarrollarse utilizando variables básicas disponibles en atención primaria, indicando potencial para mejoras adicionales manteniendo el balance entre rendimiento e interpretabilidad.