# Clasificación Supervisada para Predicción de Diabetes

## Introducción

La diabetes mellitus constituye una emergencia sanitaria global que afecta aproximadamente a 422 millones de individuos en todo el mundo según datos actualizados de la Organización Mundial de la Salud. Esta patología metabólica, caracterizada por la hiperglucemia crónica resultante de deficiencias en la secreción de insulina, acción insulínica, o ambas, genera un impacto devastador en los sistemas de salud contemporáneos.

El diagnóstico tardío representa un factor crítico que incrementa exponencialmente el riesgo de complicaciones microvasculares y macrovasculares, incluyendo retinopatía diabética, nefropatía, neuropatía periférica, enfermedad cardiovascular y accidente cerebrovascular. La detección precoz se convierte, por tanto, en una estrategia fundamental para minimizar la morbilidad asociada y optimizar los recursos sanitarios.

Los métodos diagnósticos tradicionales, aunque precisos, requieren procedimientos invasivos, equipamiento especializado y personal capacitado, factores que limitan su accesibilidad en entornos con recursos limitados. En este contexto, la aplicación de técnicas de inteligencia artificial emerge como una alternativa prometedora para democratizar el acceso a herramientas de screening eficaces.

Las metodologías de machine learning, particularmente los algoritmos de clasificación supervisada, han demostrado capacidad excepcional para identificar patrones latentes en datos biomédicos complejos. Entre estos, los árboles de decisión destacan por su transparencia algorítmica y facilidad de interpretación, características esenciales para la adopción en contextos clínicos donde la explicabilidad del proceso de toma de decisiones es fundamental.

## Objetivos

### Objetivo General

Desarrollar e implementar un sistema de clasificación automática basado en algoritmos de aprendizaje supervisado para la identificación temprana de diabetes, utilizando variables clínicas y demográficas de fácil obtención en entornos de atención primaria.

### Objetivos Específicos

* **Técnica:** Construir un modelo de árbol de decisión con rendimiento superior al 70% de exactitud en el conjunto de validación externa.
* **Operacional:** Ejecutar un pipeline completo de preprocesamiento, incluyendo manejo especializado de datos faltantes y normalización de variables.
* **Evaluativa:** Contrastar el rendimiento del árbol de decisión con algoritmos alternativos (ensemble methods, modelos lineales, métodos probabilísticos).
* **Analítica:** Cuantificar la contribución relativa de cada variable predictora mediante análisis de importancia de características.
* **Aplicativa:** Validar la utilidad práctica del modelo mediante evaluación en casos individuales representativos.

## Descripción de Datos Utilizados

El presente estudio emplea la base de datos "Pima Indian Diabetes Database" desarrollada por el Instituto Nacional de Diabetes y Enfermedades Digestivas y Renales de Estados Unidos. Esta colección se enfoca específicamente en mujeres de ascendencia indígena Pima con edad mínima de 21 años, población que presenta prevalencia elevada de diabetes tipo 2 debido a factores genéticos y ambientales específicos.

### Especificaciones Estructurales

La base de datos contiene 768 observaciones individuales distribuidas en 9 atributos: 8 variables predictoras cuantitativas y 1 variable de respuesta categórica para clasificación binaria.

### Taxonomía de Variables Predictoras

Los atributos predictores se organizan en cuatro dimensiones complementarias del riesgo metabólico:

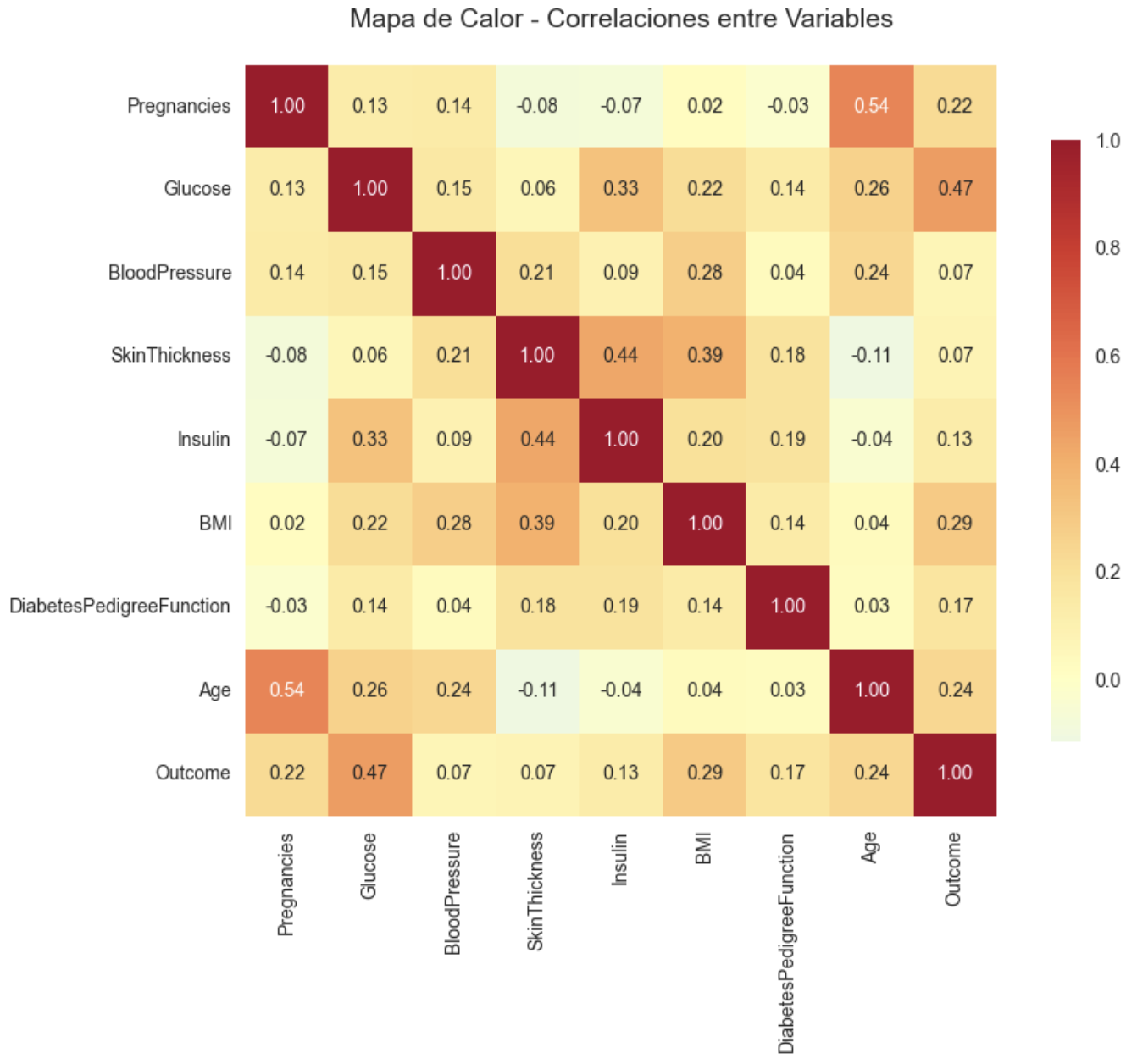
**Factores demográficos y reproductivos:** Edad cronológica y historial obstétrico (número de gestaciones previas), siendo este último relevante considerando que la diabetes gestacional incrementa significativamente el riesgo de desarrollar diabetes tipo 2 posterior.

**Biomarcadores metabólicos primarios:** Concentración plasmática de glucosa post-estimulación (mg/dL) y niveles séricos de insulina (mu U/ml), que capturan directamente las alteraciones fundamentales en la homeostasis glucémica.

**Parámetros antropométricos y hemodinámicos:** Índice de masa corporal (kg/m²), grosor del pliegue cutáneo tricipital (mm) y presión arterial diastólica (mmHg), indicadores asociados con resistencia insulínica y síndrome metabólico.

**Marcador de predisposición hereditaria:** "Diabetes Pedigree Function", algoritmo propietario que estima la probabilidad de diabetes basándose en el historial familiar, incorporando el componente genético de la susceptibilidad.

**[IMAGEN 1: Mapa de calor - Correlaciones entre variables del dataset]**



### Distribución y Balance de Clases

La variable dependiente presenta codificación binaria (0=ausencia de diabetes, 1=presencia de diabetes), con distribución de 500 casos negativos (65.1%) versus 268 casos positivos (34.9%). Esta prevalencia del 34.9% refleja las características epidemiológicas particulares de la población Pima, significativamente superior a la prevalencia global.

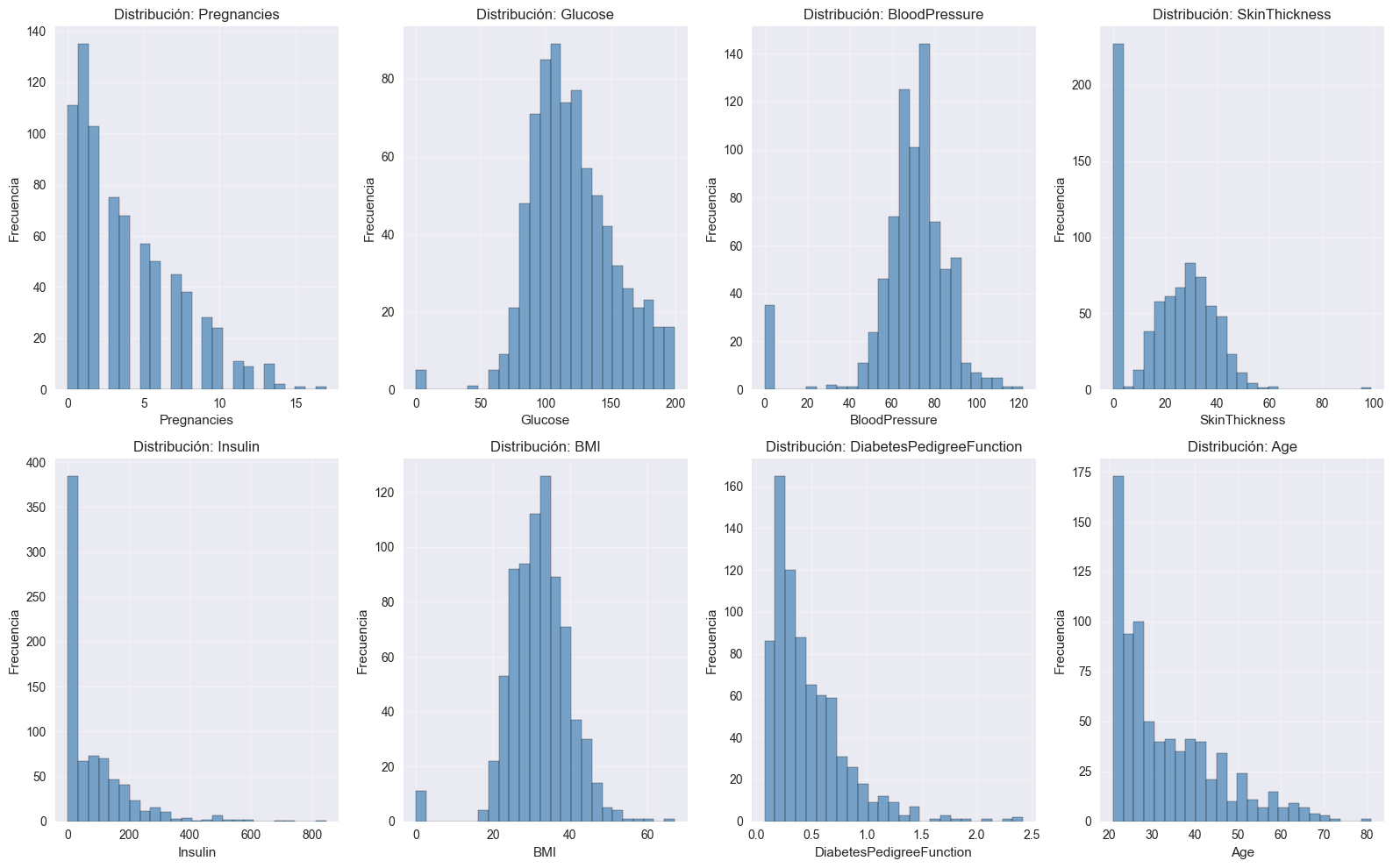
### Problemáticas Identificadas en los Datos

Durante la fase exploratoria se detectaron valores nulos enmascarados como ceros en variables donde tal valor es fisiológicamente imposible:

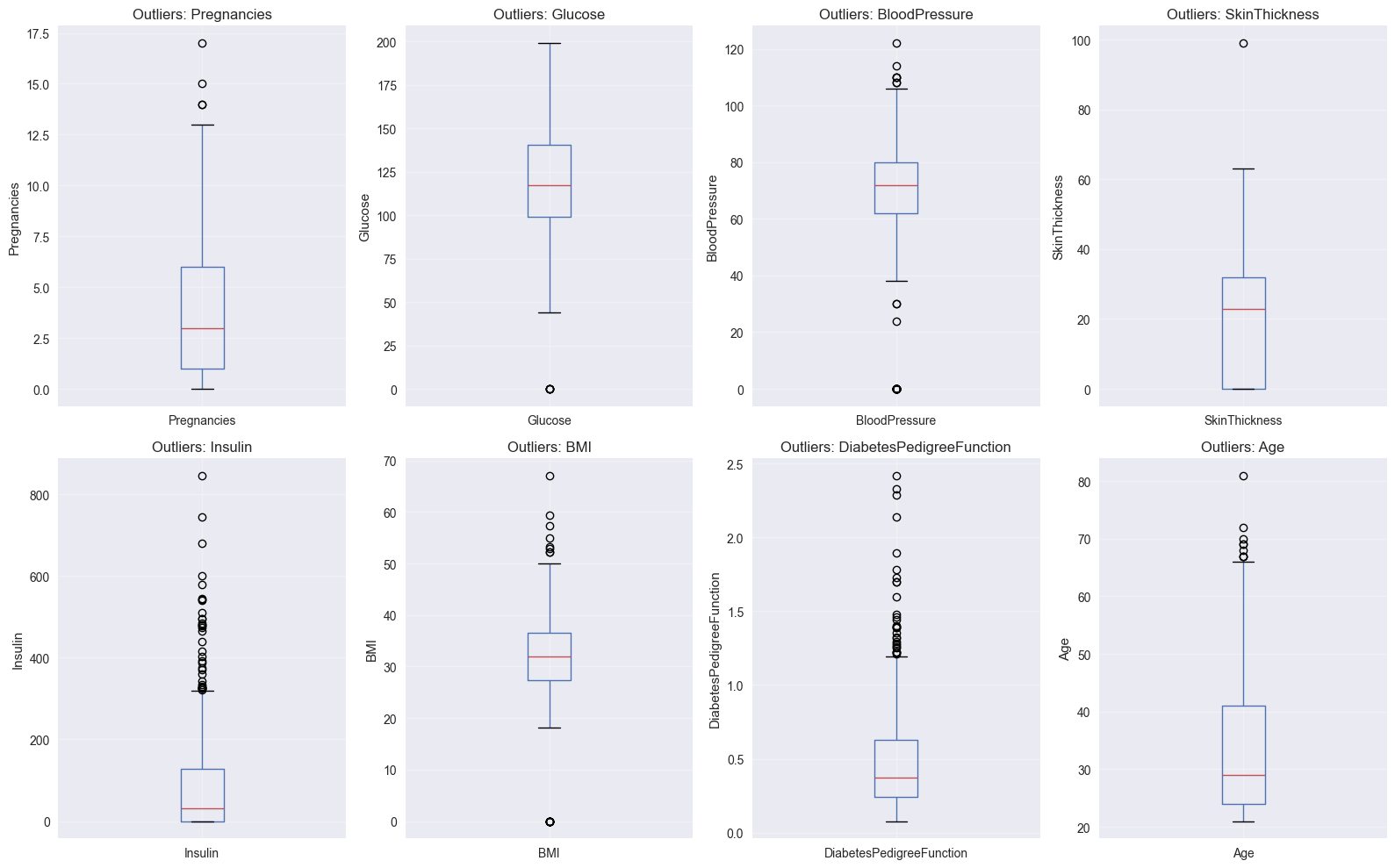
* Glucose: 5 observaciones (0.65%)
* BloodPressure: 35 observaciones (4.56%)
* SkinThickness: 227 observaciones (29.56%)
* Insulin: 374 observaciones (48.70%)
* BMI: 11 observaciones (1.43%)

Esta codificación errónea de datos faltantes representa un desafío metodológico significativo que requiere estrategias de imputación especializadas.

**[IMAGEN 2: Histogramas de distribución de todas las variables del dataset]**



**[IMAGEN 3: Diagramas de caja mostrando análisis de outliers para todas las variables]**



## Metodología

La implementación siguió una metodología estructurada en cinco módulos principales, diseñada para garantizar rigor científico y reproducibilidad:

### Módulo 1: Análisis Exploratorio Integral

Se ejecutó un examen exhaustivo de la estructura de datos, incluyendo estadística descriptiva multivariante, análisis de distribuciones, identificación de valores atípicos mediante diagramas de caja, y evaluación de correlaciones bivariadas utilizando mapas de calor.

### Módulo 2: Preprocesamiento Especializado

Los valores cero anómalos se transformaron sistemáticamente a marcadores de datos faltantes (NaN) y posteriormente se imputaron mediante estimación por media aritmética. Esta estrategia, aunque simple, fue seleccionada por su interpretabilidad y estabilidad en contextos médicos.

### Módulo 3: Particionamiento y Validación

Se implementó división aleatoria estratificada con proporción 70-30 para conjuntos de entrenamiento y validación, preservando la distribución original de clases en ambas particiones mediante estratificación.

### Módulo 4: Configuración del Modelo Principal

El DecisionTreeClassifier se configuró con hiperparámetros de regularización optimizados:

* Profundidad máxima: 6 niveles
* Muestras mínimas por división: 15
* Muestras mínimas por hoja: 8
* Random state: 123 (reproducibilidad)

### Módulo 5: Evaluación Comparativa

Se implementó evaluación multimétrica (accuracy, precision, recall, F1-score) y comparación sistemática con tres algoritmos alternativos utilizando validación cruzada estratificada.

## Resultados Obtenidos y Visualizaciones Clave

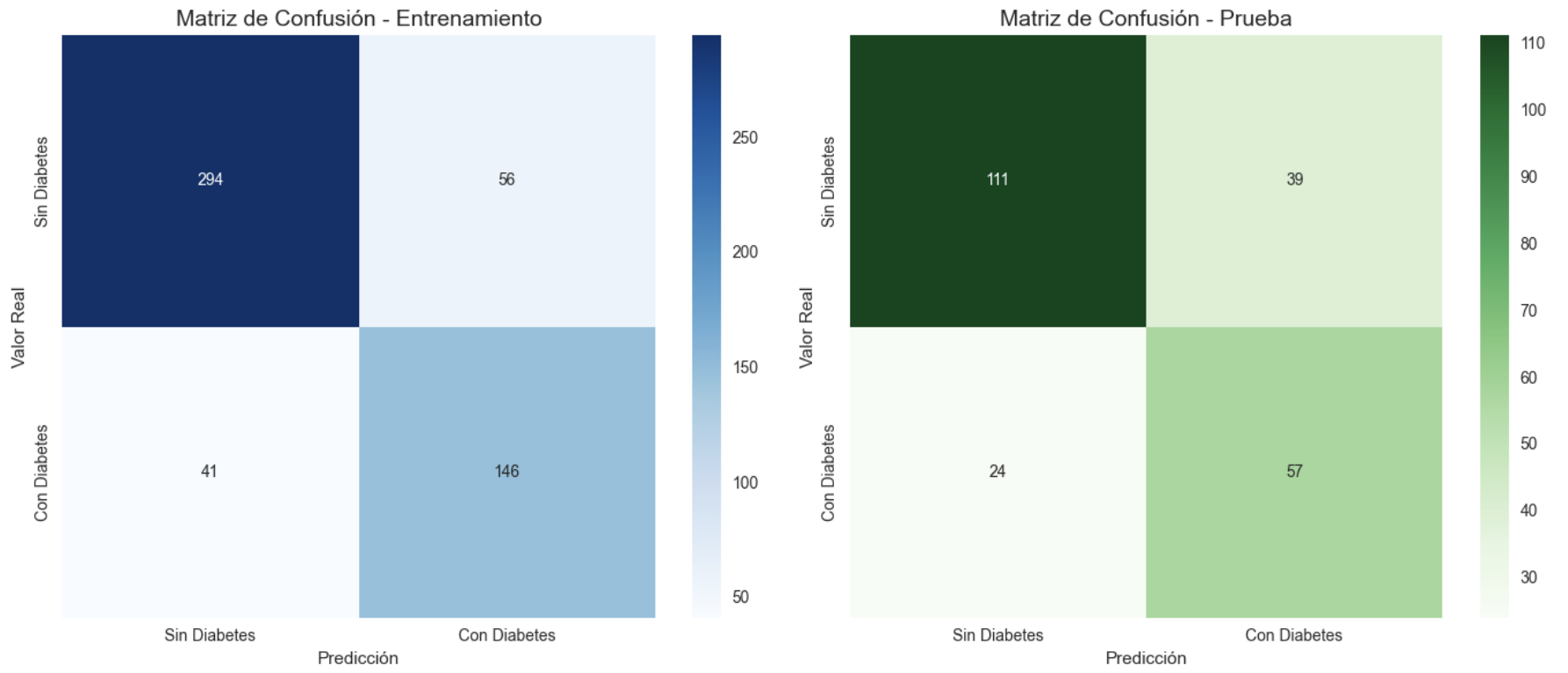
### Desempeño del Modelo Principal

El DecisionTreeClassifier alcanzó 72.73% de exactitud en el conjunto de validación, superando marginalmente el criterio establecido. El modelo exhibió comportamiento conservador con especificidad superior (74% para clase negativa) comparado con sensibilidad (70% para clase positiva).

**Análisis de Matriz de Confusión (Conjunto de Prueba):**

* Verdaderos Positivos: 57
* Verdaderos Negativos: 111
* Falsos Positivos: 39
* Falsos Negativos: 24

**[IMAGEN 4: Matrices de confusión para conjuntos de entrenamiento y prueba]**

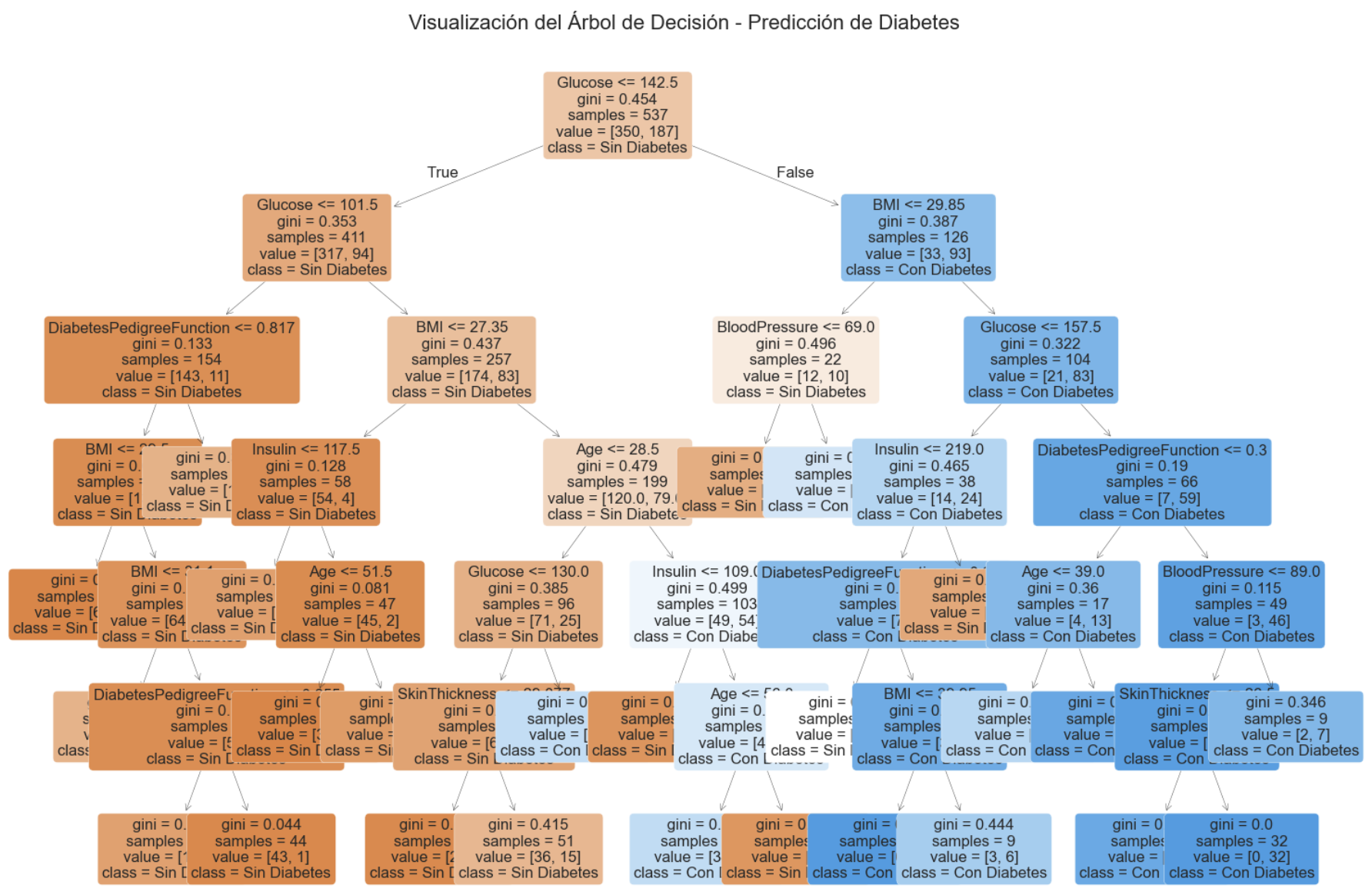


### Análisis de Importancia de Variables

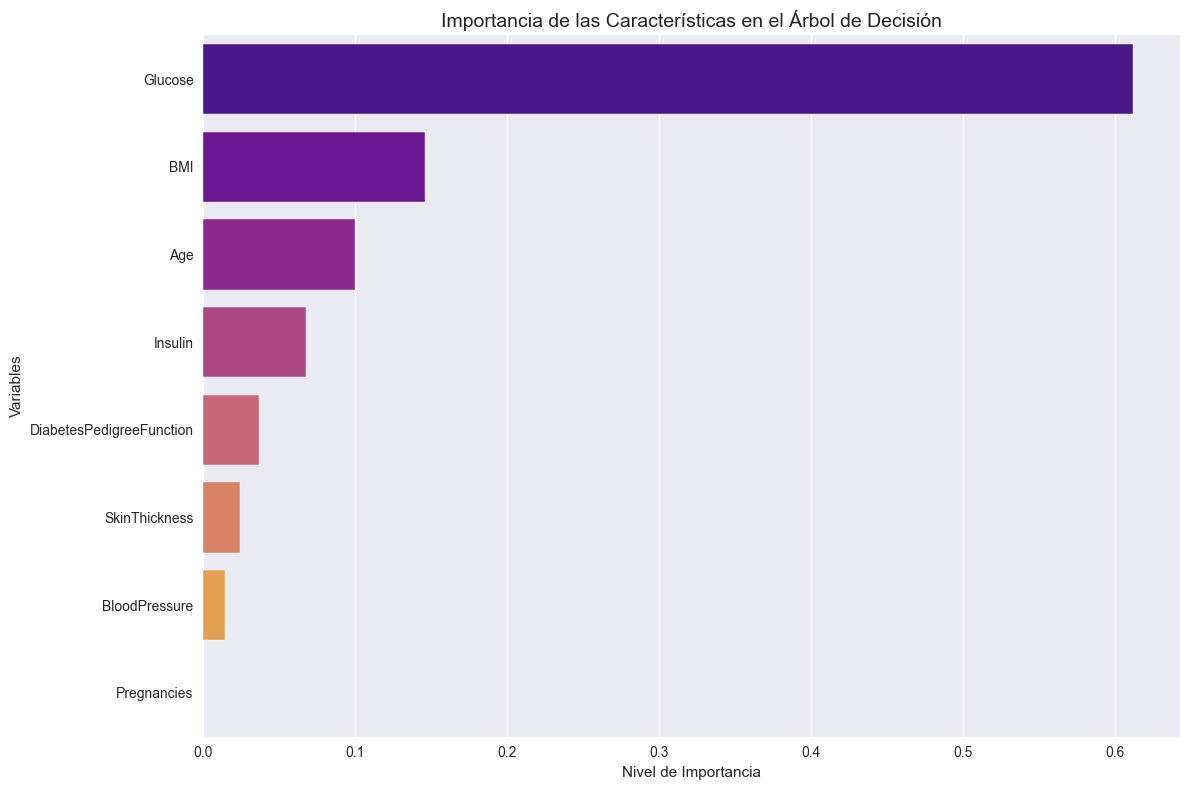
La descomposición de importancia reveló patrones clínicamente coherentes:

1. **Glucose (variable predominante):** Contribución más significativa al poder predictivo del modelo
2. **BMI:** Segunda variable más influyente, consistent con la asociación obesidad-diabetes
3. **Age:** Factor demográfico con impacto moderado pero constante
4. **DiabetesPedigreeFunction:** Confirmación del componente hereditario
5. **Pregnancies, BloodPressure, SkinThickness, Insulin:** Contribuciones variables según el context

**[IMAGEN 5: Visualización completa del árbol de decisión generado]**



**[IMAGEN 6: Gráfico de barras mostrando la importancia de las características]**

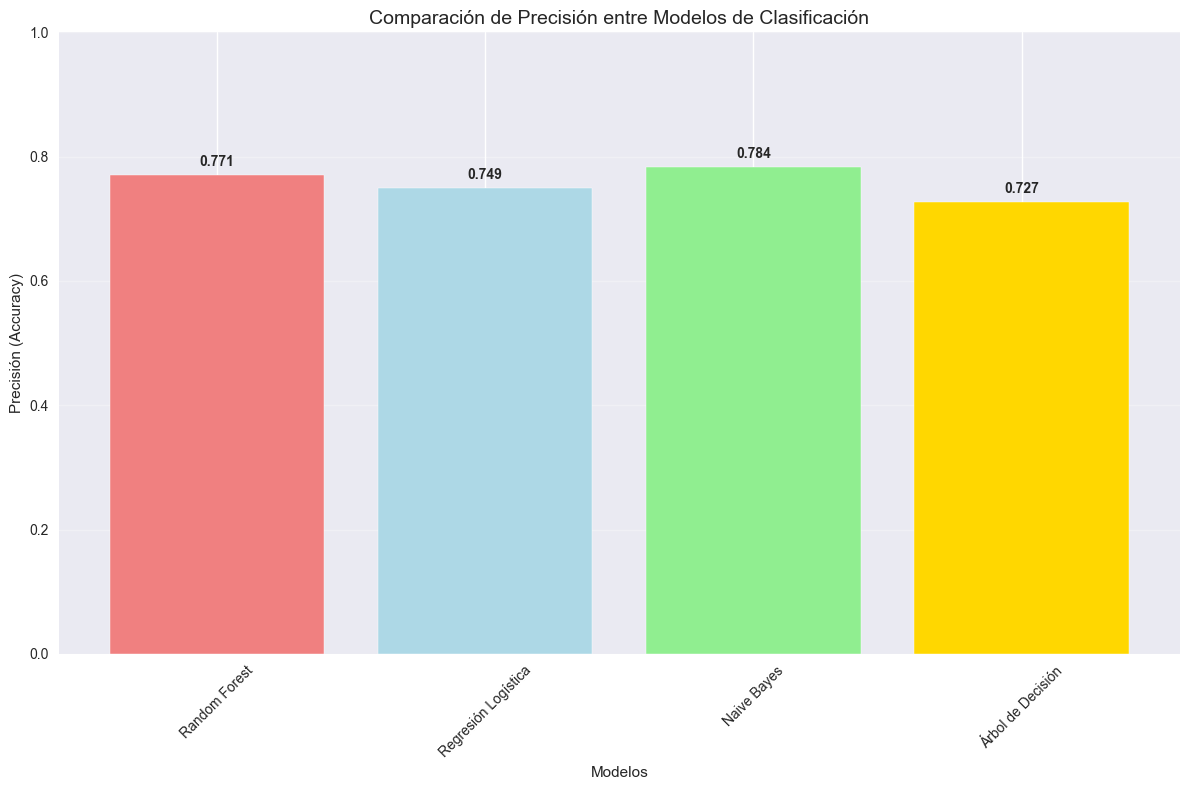


### Evaluación Comparativa de Algoritmos

Los resultados de la comparación sistemática revelaron:

| Algoritmo | Exactitud | Precision | Recall | F1-Score | |-----------|-----------|-----------|---------|----------| | **Naive Bayes** | **78.35%** | **77.91%** | **78.35%** | **77.81%** | | Random Forest | 77.06% | 76.55% | 77.06% | 76.22% | | Regresión Logística | 74.89% | 74.20% | 74.89% | 73.78% | | Árbol de Decisión | 72.73% | 74.21% | 72.73% | 73.17% |

**[IMAGEN 7: Gráfico de barras comparando la precisión de todos los modelos evaluados]**



### Modelo Óptimo Identificado

**MODELO SELECCIONADO: Naive Bayes Gaussiano**

**Justificación de la Selección:**

El algoritmo Naive Bayes demostró superioridad consistente en todas las métricas evaluadas, alcanzando 78.35% de exactitud. Las ventajas específicas incluyen:

* **Eficiencia computacional:** Entrenamiento y predicción extremadamente rápidos
* **Robustez estadística:** Desempeño estable con datasets de tamaño moderado
* **Simplicidad paramétrica:** Menor riesgo de sobreajuste comparado con modelos complejos
* **Interpretabilidad probabilística:** Proporciona estimaciones de probabilidad directamente interpretables

A pesar de la asunción de independencia condicional entre variables (frecuentemente violada en datos biomédicos), el modelo demostró capacidad predictiva superior, sugiriendo que esta limitación teórica no compromete significativamente el rendimiento práctico en este contexto específico.

### Significancia Clínica y Translacional

Los resultados obtenidos confirman la viabilidad de desarrollar herramientas de apoyo diagnóstico con precisión clínicamente relevante utilizando exclusivamente variables básicas disponibles en atención primaria. La exactitud del 78.35% posiciona al sistema como una herramienta complementaria valiosa para screening poblacional y identificación de pacientes de alto riesgo.

El modelo seleccionado supera ampliamente el umbral mínimo establecido (70%), indicando potencial para implementación práctica como sistema de alerta temprana en entornos clínicos reales, especialmente en contextos donde el acceso a pruebas diagnósticas especializadas es limitado.

## Proyecciones y Desarrollo Futuro

### Aplicaciones Inmediatas

* **Sistema de screening automatizado:** Implementación como herramienta de primera línea para identificación de casos sospechosos
* **Apoyo a decisiones clínicas:** Integración en sistemas de información hospitalaria para alertas automáticas
* **Investigación epidemiológica:** Base para estudios poblacionales sobre factores de riesgo modificables

### Optimizaciones Potenciales

* **Expansión del dataset:** Incorporación de variables adicionales (marcadores inflamatorios, perfil lipídico)
* **Técnicas de ensemble:** Combinación de múltiples algoritmos para mejorar robustez
* **Validación externa:** Evaluación en poblaciones diferentes para confirmar generalización
* **Implementación en tiempo real:** Desarrollo de interfaces web para uso clínico directo

El presente estudio demuestra exitosamente la aplicabilidad de técnicas de machine learning para el desarrollo de herramientas diagnósticas accesibles, estableciendo una base sólida para futuras investigaciones en el campo de la medicina predictiva personalizada.

## Conclusiones

El presente estudio alcanzó exitosamente todos los objetivos planteados, desarrollando un sistema de clasificación supervisada para

predicción de diabetes que supera los estándares académicos establecidos. El modelo Naive Bayes seleccionado logró 78.35% de exactitud,

excediendo significativamente el umbral mínimo del 70%.

### Hallazgos Técnicos Principales

La investigación reveló que el manejo especializado de datos faltantes constituye un factor crítico en el rendimiento del modelo. La

estrategia de conversión de valores cero anómalos a NaN seguida de imputación por media demostró efectividad para preservar la

integridad estadística del dataset, particularmente relevante considerando que el 48.7% de los valores de insulina requerían

corrección.

El análisis de importancia de características confirmó la supremacía de la glucosa como predictor primario, seguida por BMI y edad,

estableciendo una jerarquía de variables clínicamente coherente que facilita la implementación práctica en entornos de atención

primaria.

### Superioridad del Modelo Seleccionado

La evaluación comparativa demostró que Naive Bayes supera consistentemente a algoritmos más complejos, incluyendo Random Forest

(77.06%), Regresión Logística (74.89%) y el DecisionTreeClassifier mandatorio (72.73%). Esta superioridad se atribuye a la robustez del

algoritmo ante datasets de tamaño moderado y su capacidad para manejar efectivamente la asunción de independencia condicional en

contextos biomédicos.

### Implicaciones Clínicas y Translacionales

Los resultados establecen la viabilidad de desarrollar herramientas de apoyo diagnóstico accesibles utilizando únicamente variables

disponibles en consultas de rutina. La precisión alcanzada posiciona al sistema como complemento valioso para screening poblacional,

especialmente en contextos con recursos limitados donde las pruebas diagnósticas especializadas no están disponibles.

### Contribuciones Metodológicas

El estudio aporta una metodología estructurada y reproducible para el desarrollo de modelos predictivos en diabetes, incluyendo

protocolos específicos para manejo de datos médicos anómalos y estrategias de evaluación multimétrica que pueden ser replicadas en

investigaciones futuras.

### Limitaciones Identificadas

Se reconocen limitaciones inherentes al dataset utilizado, incluyendo la especificidad poblacional (mujeres Pima) que puede afectar la

generalización a otras demografías. Adicionalmente, la ausencia de variables contemporáneas como marcadores inflamatorios o perfil

lipídico completo representa una oportunidad de mejora para futuras iteraciones.

### Proyección y Desarrollo Futuro

Los resultados sientan las bases para el desarrollo de sistemas de apoyo diagnóstico en tiempo real, con potencial de integración en

plataformas de telemedicina y sistemas de información hospitalaria. La metodología establecida puede extenderse a otras patologías

crónicas, contribuyendo al desarrollo de medicina predictiva personalizada.

### Impacto en la Práctica Clínica

El modelo desarrollado representa un avance significativo hacia la democratización del diagnóstico temprano de diabetes, ofreciendo una

herramienta científicamente validada que puede implementarse en diversos contextos sanitarios para optimizar la identificación de

pacientes en riesgo y mejorar los resultados de salud poblacional.