# Reporte: Modelo Random Forest para Clasificación de Diabetes

## Justificación del Algoritmo

El algoritmo Random Forest fue seleccionado por su capacidad para manejar conjuntos de datos con múltiples características, incluso si algunas no son altamente predictivas. Es robusto frente al sobreajuste y proporciona una métrica de importancia de características, lo cual es útil para entender qué variables son más relevantes en la predicción de diabetes.

## Descripción del Diseño del Modelo

El modelo fue diseñado siguiendo los pasos descritos a continuación:

1. Carga y preprocesamiento de datos: Se cargaron los datos del archivo 'diabetes.csv' y se inspeccionaron para identificar valores nulos y estadísticas básicas. No fue necesario realizar un tratamiento especial, ya que los datos estaban completos.

2. División del conjunto de datos: Los datos fueron divididos en conjuntos de entrenamiento (80%) y prueba (20%) de forma estratificada para garantizar una representación equitativa de ambas clases.

3. Entrenamiento del modelo base: Se entrenó un modelo Random Forest inicial con 100 estimadores y otros hiperparámetros predeterminados para establecer una línea base de precisión.

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

## Explicación de la Optimización

Para mejorar la precisión del modelo, se optimizaron los hiperparámetros utilizando Random Search y Grid Search. Random Search permitió explorar un rango amplio de combinaciones de hiperparámetros, mientras que Grid Search afinó los resultados alrededor de los mejores valores obtenidos. Los hiperparámetros ajustados incluyeron el número de estimadores, la profundidad máxima del árbol, el número mínimo de muestras por nodo y el criterio de calidad de división ('gini' o 'entropy').

El mejor conjunto de hiperparámetros encontrado fue:  
- Número de estimadores: 64  
- Profundidad máxima: 15  
- Mínimo de muestras para dividir un nodo: 8  
- Mínimo de muestras en una hoja: 1  
- Criterio de calidad: 'gini'  
Con esta configuración, el modelo alcanzó una precisión del 78.34% en el conjunto de prueba.

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

## Gráfica Personalizada e Interpretación de Resultados

En esta sección se deben incluir las gráficas generadas en el entorno Jupyter Notebook:

1. Matriz de Confusión: Permite visualizar los aciertos y errores del modelo en ambas clases (diabetes y no diabetes). Se deben incluir las matrices para el modelo base y el optimizado.

2. Curva ROC: Muestra la capacidad de discriminación del modelo. Se incluye la curva para el modelo base y el optimizado, indicando el área bajo la curva (AUC) para cada uno.

Gráfico

Descripción generada automáticamente

Gráfico, Gráfico de líneas

Descripción generada automáticamente