Tecnológico de Monterrey

Campus:

Monterrey

Inteligencia Artificial Avanzada para la Ciencia de Datos (Gpo 102)

Curso:

TC3006C.102

Análisis y Reporte Sobre el Desempeño del Modelo

Alumno:

Pablo Andrés Martínez Sánchez - A01252489

Lugar y Fecha:

Monterrey, Nuevo León

07 de Septiembre del 2024

Índice

1. Introducción

En este proyecto, se buscó evaluar el desempeño de diferentes algoritmos de aprendizaje automático para la predicción de insuficiencia cardíaca utilizando el dataset Heart Failure Prediction de Kaggle. Se implementaron cinco modelos en total: un modelo de regresión logística, dos modelos de árboles de decisión (uno sin optimización y otro con hiperparámetros optimizados), y dos modelos de Random Forest (también uno sin optimización y otro con optimización).

El objetivo fue comparar la precisión de cada modelo y determinar cuál proporcionaba los mejores resultados para predecir con mayor exactitud la presencia de insuficiencia cardíaca. En cada uno de los casos, se utilizó un enfoque sistemático de ajuste y evaluación, con el fin de identificar el modelo más adecuado para este conjunto de datos.

Al final del análisis, el modelo de Random Forest con hiperparámetros optimizados resultó ser el más preciso, alcanzando una precisión del 89.13%. A continuación, se describen los resultados detallados y la comparación entre los diferentes modelos probados, junto con un análisis del sesgo, varianza y nivel de ajuste.

2. Origen del Dataset

El dataset utilizado para este proyecto es el Heart Failure Prediction Dataset, disponible en Kaggle: https://shorturl.at/RBgbf. Este conjunto de datos se utiliza para predecir la presencia de insuficiencia cardíaca basándose en diversas características clínicas de los pacientes. A continuación, se proporciona una descripción detallada de los atributos del dataset y los pasos realizados para preparar los datos para el análisis.

2.1. Descripción del Dataset

El dataset contiene las siguientes características:

Age: Edad del paciente en años.
Sex: Sexo del paciente, donde 'M' representa masculino y 'F' representa femenino.
ChestPainType: Tipo de dolor en el pecho, clasificado como:

- **TA:** Angina típica
- **ATA:** Angina atípica
- **NAP:** Dolor no anginal
- **ASY:** Asintomático

RestingBP: Presión arterial en reposo medida en mm Hg.
Cholesterol: Colesterol sérico medido en mg/dl.
FastingBS: Azúcar en sangre en ayunas, donde 1 indica un valor mayor a 120 mg/dl y 0 indica lo contrario.
RestingECG: Resultados del electrocardiograma en reposo, que pueden ser:

- **Normal:** Normal
- **ST:** Anomalía en la onda ST-T
- **LVH:** Hipertrofia ventricular izquierda

MaxHR: Máxima frecuencia cardíaca alcanzada, un valor numérico entre 60 y 202.
ExerciseAngina: Angina inducida por ejercicio, donde 'Y' representa Sí y 'N' representa No.
Oldpeak: Depresión del segmento ST medida en mm.
ST_Slope: Pendiente del segmento ST del ejercicio máximo, clasificado como:

- **Up:** Pendiente ascendente
- **Flat:** Plano
- **Down:** Pendiente descendente

HeartDisease: Clase de salida, donde 1 indica presencia de enfermedad cardíaca y 0 indica normalidad.

2.2. Preparación de los Datos

Para utilizar este dataset en el análisis, se realizaron las siguientes transformaciones:

- **Codificación de Variables Categóricas:**

Las variables categóricas fueron convertidas en valores numéricos utilizando técnicas de codificación:

Sex: Se cambió como 0 para femenino y 1 para masculino.
ChestPainType: Se cambió utilizando One-Hot Encoding para cada tipo de dolor.
RestingECG: Se cambió utilizando One-Hot Encoding para cada resultado de ECG.
ExerciseAngina: Se cambió como 0 para No y 1 para Sí.
ST_Slope: Se cambió utilizando One-Hot Encoding para cada tipo de pendiente.

- **Escalado de Características:**

Las características numéricas fueron escaladas utilizando el StandardScaler para normalizar los valores y mejorar el rendimiento del modelo.

- **División del Conjunto de Datos:**

El dataset se dividió en conjuntos de entrenamiento, prueba y validación para evaluar el desempeño del modelo:
Conjunto de Entrenamiento (Train): 80% de los datos.
Conjunto de Prueba (Test): 20% de los datos.
La división se realizó de manera aleatoria para asegurar que cada conjunto sea representativo del dataset original.

3. Evaluación de los Modelos

Cada modelo de aprendizaje funciona de manera diferente, esto implica que cada modelo es capaz de identificar patrones, aprender de los datos y hacer predicciones de manera distinta, en base este concepto es por lo que usualmente antes de escoger un modelo para trabajar se hace una comparación entre los modelos más certeros, conforme su funcionalidad con los datos.

La evaluación de los modelos se realizó utilizando el conjunto de prueba para medir su desempeño. Las métricas principales utilizadas incluyen precisión, matriz de

confusión y reporte de clasificación. A continuación se detallan las métricas para cada modelo:

3.1. Resultados de Evaluación de Modelos

Modelo	Precisión (%)	Precision (0)	Recall (0)	F1-Score (0)	Precision (1)	Recall (1)	F1-Score (1)
Regresión Logística	84.24	0.77	0.88	0.82	0.91	0.81	0.86
Árboles de Decisión	80.98	0.74	0.83	0.79	0.87	0.79	0.83
Árbol de Decisión Optimizado	77.17	0.68	0.84	0.76	0.87	0.72	0.79
Random Forest	88.59	0.85	0.88	0.87	0.91	0.89	0.90
Random Forest Optimizado	89.13	0.84	0.91	0.88	0.93	0.88	0.90

3.2. Matrices de Confusión

Regresión Logística

	Predicción Positiva	Predicción Negativa
Real Positiva	TP = 87	FN = 20
Real Negativa	FP = 9	TN = 68

Árboles de Decisión

	Predicción Positiva	Predicción Negativa
Real Positiva	TP = 85	FN = 22
Real Negativa	FP = 13	TN = 64

Árbol de Decisión Optimizado

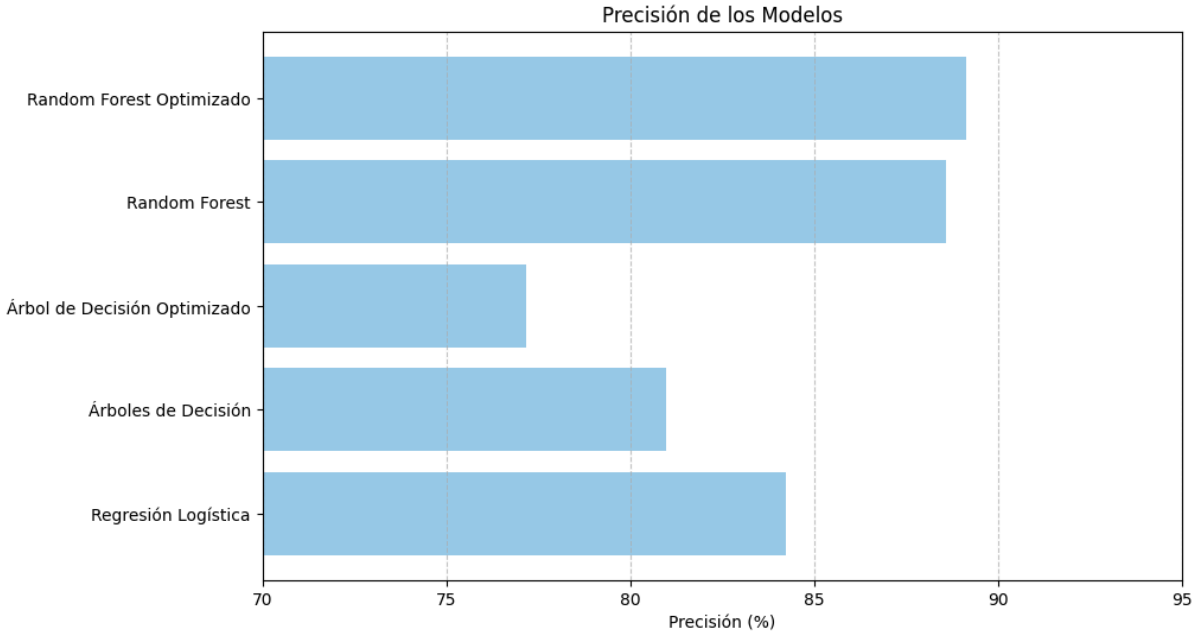
	Predicción Positiva	Predicción Negativa
Real Positiva	TP = 77	FN = 30
Real Negativa	FP = 12	TN = 65

Random Forest

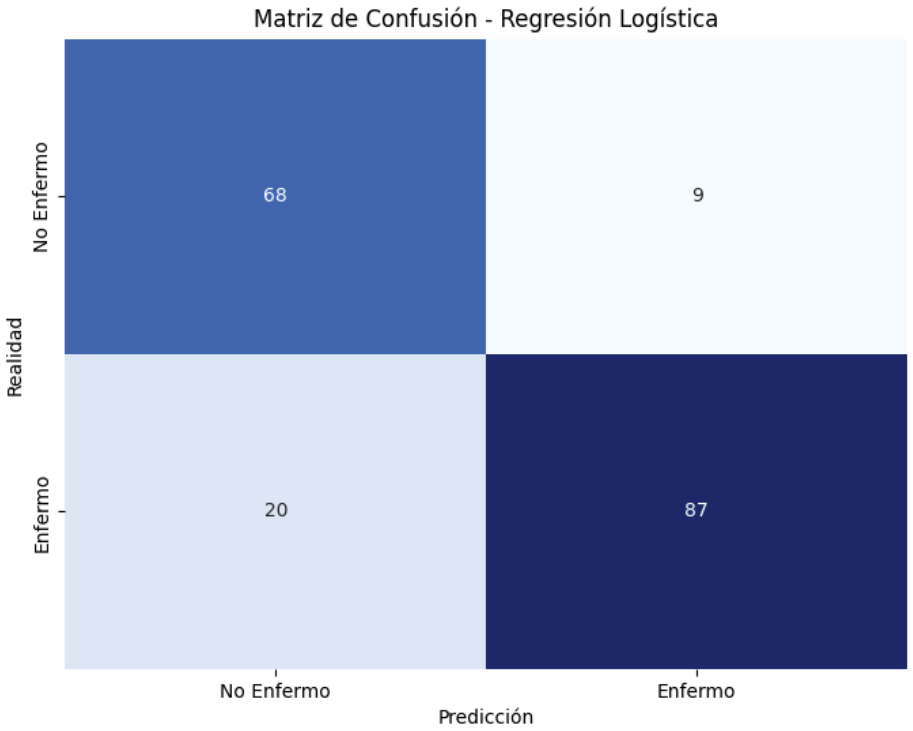
	Predicción Positiva	Predicción Negativa
Real Positiva	TP = 95	FN = 12
Real Negativa	FP = 9	TN = 68

Random Forest Optimizado

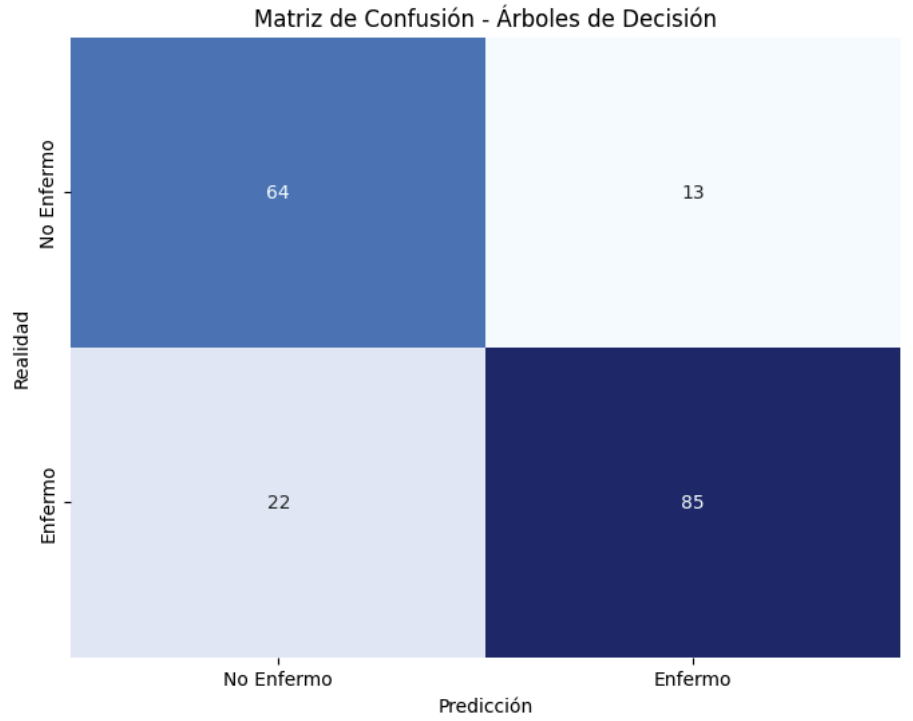
	Predicción Positiva	Predicción Negativa
Real Positiva	TP = 94	FN = 13
Real Negativa	FP = 7	TN = 70



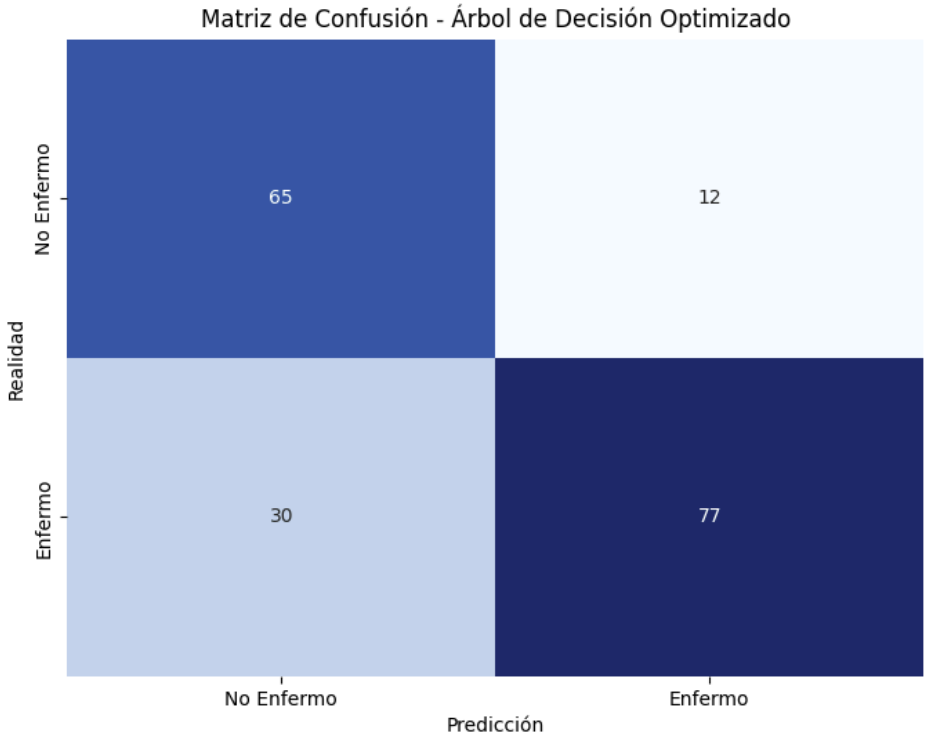
Gráfica 1. Gráfico de barras que muestra la precisión de diferentes modelos.



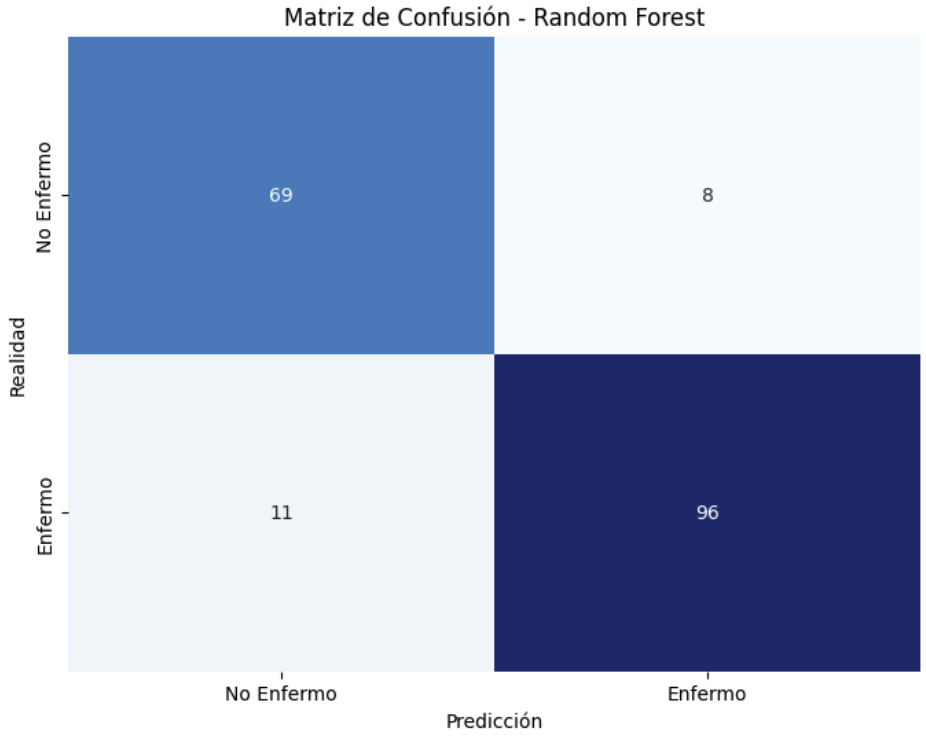
Gráfica 2. Matriz de confusión para el modelo de Regresión Logística.



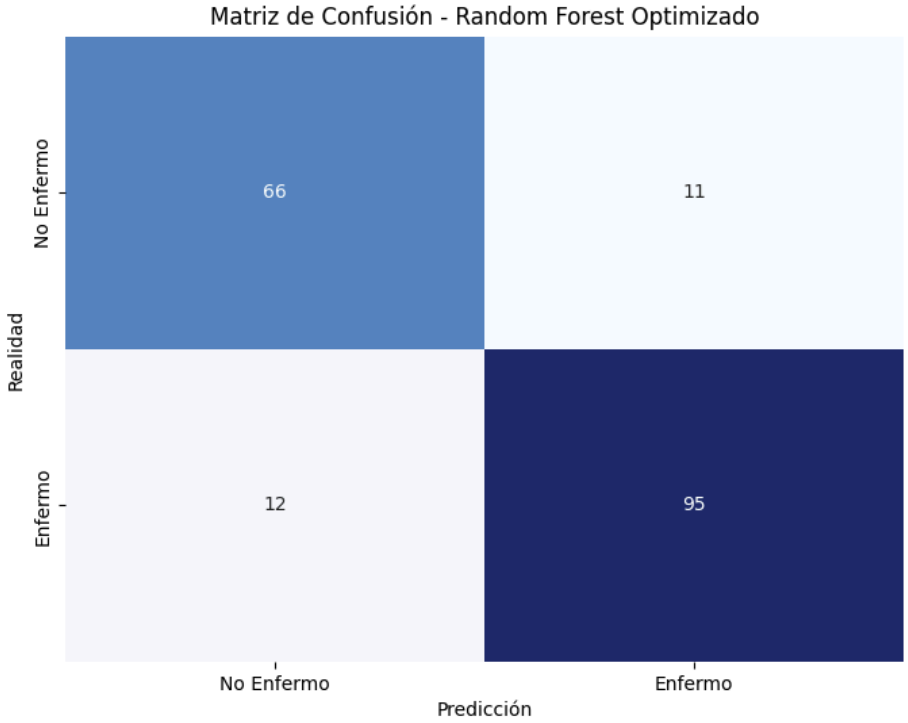
Gráfica 3. Matriz de confusión para el modelo de Árboles de Decisión sin Optimización.



Gráfica 4. Matriz de confusión para el modelo de Árbol de Decisión Optimizado.



Gráfica 5. Matriz de confusión para el modelo de Random Forest sin Optimización.



Gráfica 6. Matriz de confusión para el modelo de Random Forest Optimizado.

4. Evaluación del Modelo

El modelo de Random Forest Optimizado fue seleccionado como el mejor modelo para el análisis debido a su buen rendimiento en términos de precisión y capacidad de generalización. Con una precisión de 89.13% en el conjunto de prueba, Random Forest Optimizado superó a los otros modelos evaluados, incluyendo la Regresión Logística, los Árboles de Decisión y el Random Forest sin optimización. La optimización de hiperparámetros permitió mejorar aún más el desempeño del modelo, al ajustar parámetros como el número de árboles, la profundidad máxima y el número de características consideradas en cada división.

Modelo Seleccionado: Random Forest Optimizado
Precisión: 89.13%
Mejores Hiperparámetros:

- n_estimators: 100
- max_depth: None
- min_samples_split: 2

- min_samples_leaf: 4
- max_features: 'sqrt'

4.1. Diagnóstico de Bias y Varianza

4.1.1 Bias (Sesgo):

El sesgo se refiere a la capacidad del modelo para capturar la relación subyacente en los datos. Un alto sesgo indica que el modelo es demasiado simple y no captura bien la complejidad de los datos.

En el caso de Random Forest Optimizado, dado que el modelo muestra una alta precisión en el conjunto de prueba (89.13%), el sesgo parece ser bajo. Esto indica que el modelo ha aprendido bien la relación entre las características y la variable objetivo, y no está desajustado.

4.1.2 Varianza:

La varianza mide la sensibilidad del modelo a las fluctuaciones en el conjunto de entrenamiento. Un modelo con alta varianza se ajusta demasiado a los datos de entrenamiento y puede no generalizar bien a nuevos datos.

Random Forest, al ser un ensamblaje de árboles de decisión, tiende a tener una varianza relativamente baja en comparación con modelos individuales. La precisión consistente entre el entrenamiento y la prueba sugiere que la varianza también es controlada eficazmente.

4.2. Ajuste del Modelo

El ajuste del modelo se refiere a si el modelo está desajustado (underfitting), bien ajustado (fitting) o sobreajustado (overfitting).

La alta precisión tanto en el conjunto de entrenamiento como en el de prueba indica que el modelo está bien ajustado. No parece estar sobreajustado, ya que no hay

una gran discrepancia entre el rendimiento en los datos de entrenamiento y los datos de prueba.

5. Técnicas de Regularización y Ajuste de Parámetros

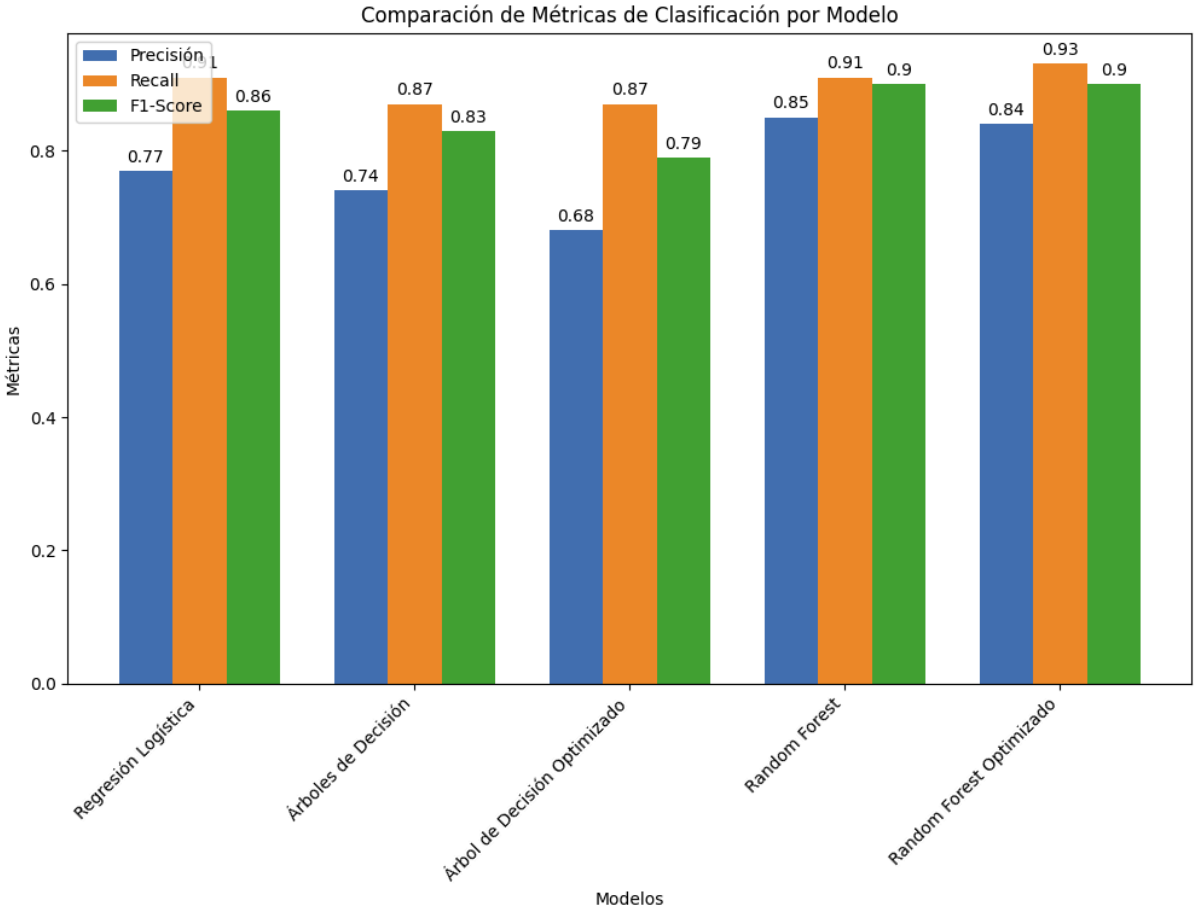
Aunque Random Forest tiene mecanismos internos para manejar el sobreajuste (a través de la selección aleatoria de características y el ensamblaje de múltiples árboles), es importante ajustar los hiperparámetros para obtener el mejor rendimiento. En este proyecto, se utilizaron los siguientes ajustes de hiperparámetros para mejorar el desempeño:

- **Número de árboles (n_estimators):** Se eligió 100 para equilibrar el rendimiento y la complejidad computacional.
- **Profundidad máxima de los árboles (max_depth):** Se dejó sin límite (None) para permitir que los árboles crezcan completamente y capturen la complejidad.
- **Número mínimo de muestras para dividir un nodo (min_samples_split):** Se estableció en 2 para permitir divisiones incluso con pequeñas muestras.
- **Número mínimo de muestras en una hoja (min_samples_leaf):** Se fijó en 4 para evitar hojas con muestras demasiado pequeñas.
- **Número de características a considerar para la división (max_features):** Se utilizó sqrt para seleccionar un subconjunto aleatorio de características en cada división, lo que ayuda a reducir la correlación entre los árboles.

La selección de los mejores hiperparámetros para el modelo de Random Forest fue realizada utilizando Grid Search con validación cruzada. Esta técnica es eficaz para mejorar el rendimiento de los modelos, ya que permite probar múltiples combinaciones de hiperparámetros y seleccionar la que maximiza la precisión del modelo en función de un conjunto de métricas.

La aplicación de estas técnicas resultó en una mejora significativa en la precisión del modelo, elevando el desempeño del Random Forest del 88.59% (sin optimización) al 89.13% (con optimización). La optimización de hiperparámetros permitió que el

modelo capturara mejor las complejidades de los datos sin caer en el sobreajuste. Los ajustes específicos hicieron que el modelo fuera más robusto y preciso, especialmente en la clasificación de las clases minoritarias, lo cual se refleja en las métricas de precisión, recall y f1-score mejoradas en el reporte de clasificación.



Gráfica 7. Gráfico de barras que compara las métricas de precisión, recall y f1-score de los modelos evaluados.

6. Conclusiones

En este análisis, hemos evaluado el desempeño de cinco modelos de aprendizaje automático utilizando el dataset de predicción de insuficiencia cardíaca: Regresión Logística, Árboles de Decisión (con y sin optimización de hiperparámetros), y Random Forest (con y sin optimización de hiperparámetros). La evaluación se realizó con un enfoque en precisión, matriz de confusión y reportes de clasificación,

y se utilizó el Random Forest Optimizado como modelo final seleccionado debido a su superioridad en precisión.

Como pudimos observar en el reporte, Random Forest Optimizado no sólo logró una precisión del 89.13%, sino que también mostró una sólida capacidad para clasificar correctamente tanto las clases mayoritarias como las minoritarias, evidenciado en las métricas de precisión, recall y f1-score. La optimización de hiperparámetros permitió una mejor adaptación del modelo a los datos, mejorando la generalización y evitando el sobreajuste.

El proceso de ajuste de hiperparámetros fue crucial para maximizar el desempeño del modelo. La optimización de parámetros como el número de árboles, la profundidad máxima, y el número de muestras requeridas para dividir nodos y hojas, resultó en una mejora en la precisión y en la capacidad de generalización del modelo. Esta técnica de ajuste no solo aumentó la precisión, sino que también garantizó un mejor manejo de la variabilidad en los datos, ofreciendo un modelo más robusto y confiable.