Álvaro González Plaza

EVALUACIÓN DE MODELOS DE CLASIFICACIÓN

En este informe haremos una comparación de clasificadores para un conjunto de datos sobre diagnósticos de cáncer de mama del **Repository of Machine Learning Databases** de UC Irvine.

Hemos utilizado el conjunto de datos **Breast Cancer Wisconsin (Diagnostic)** donado en 1995 y sujeto al área de Salud y Medicina. Este conjunto de datos incluye características calculadas a partir de una imagen digitalizada de una aspiración con aguja fina de masa mamaria.

Se realizaron las siguientes tareas:

-Preprocesamiento de datos.

-Evaluación de modelos mediante validación cruzada K-Fold(k=10).

-Ajuste de hiperparámetros.

-Generación de curvas ROC

- **Descripción del Conjunto de Datos**

-Número de muestras: 569

-Número de características: 30

-Número de clases: 2 (Benigno - 357, Maligno - 212)

Cada muestra en el conjunto de datos incluye varias características calculadas a partir de imágenes digitalizadas de biopsias de mama, como el radio, la textura, el perímetro, el área, la suavidad, la compacidad, la concavidad, los puntos cóncavos, la simetría y la dimensión fractal.

- **Preprocesamiento de los Datos**

El preprocesamiento de los datos incluyó los siguientes pasos:

-Eliminación de la columna ID.

-Conversión de la columna Diagnosis a valores binarios (1 para Maligno, 0 para Benigno).

-Escalado de características usando 'StandardScaler'.

-División de los datos en conjuntos de entrenamiento y prueba (70% - 30%).

Evaluación de Modelos

Se han implementado y evaluado a los siguientes clasificadores utilizando validación cruzada 10-fold para asegurar robustez en nuestros resultados:

-Clasificador Naive Bayes

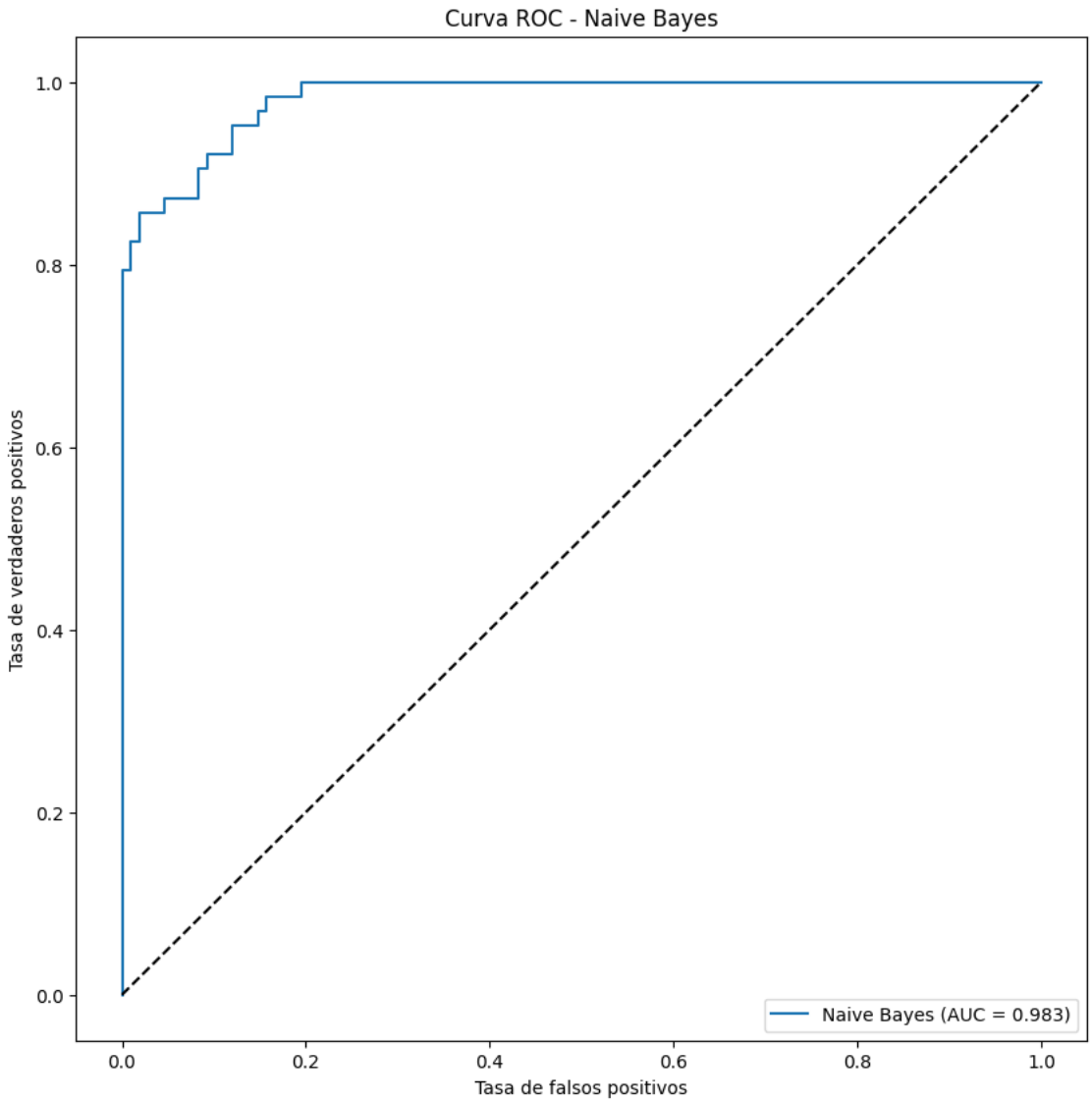
-Clasificador Nearest Neighbors

-Clasificador árbol de decisión

-Clasificador Support Vector Machine

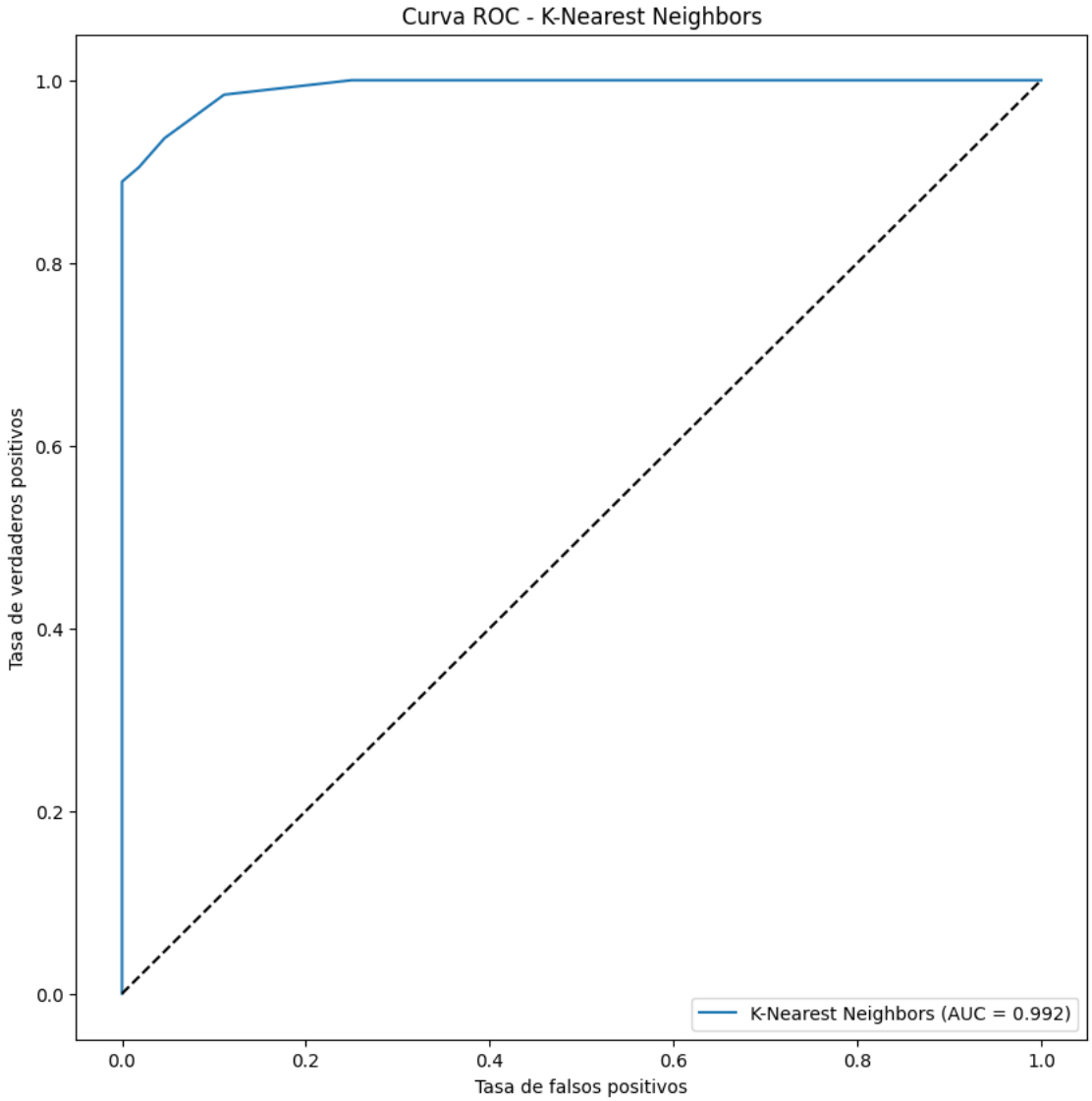
CLASIFICADOR NAIVE BAYES

El clasificador Naive Bayes (GaussianNB) no requiere ajuste de hiperparámetros. Utilizando validación cruzada 10-fold, se obtuvo una exactitud promedio de 0.9315.



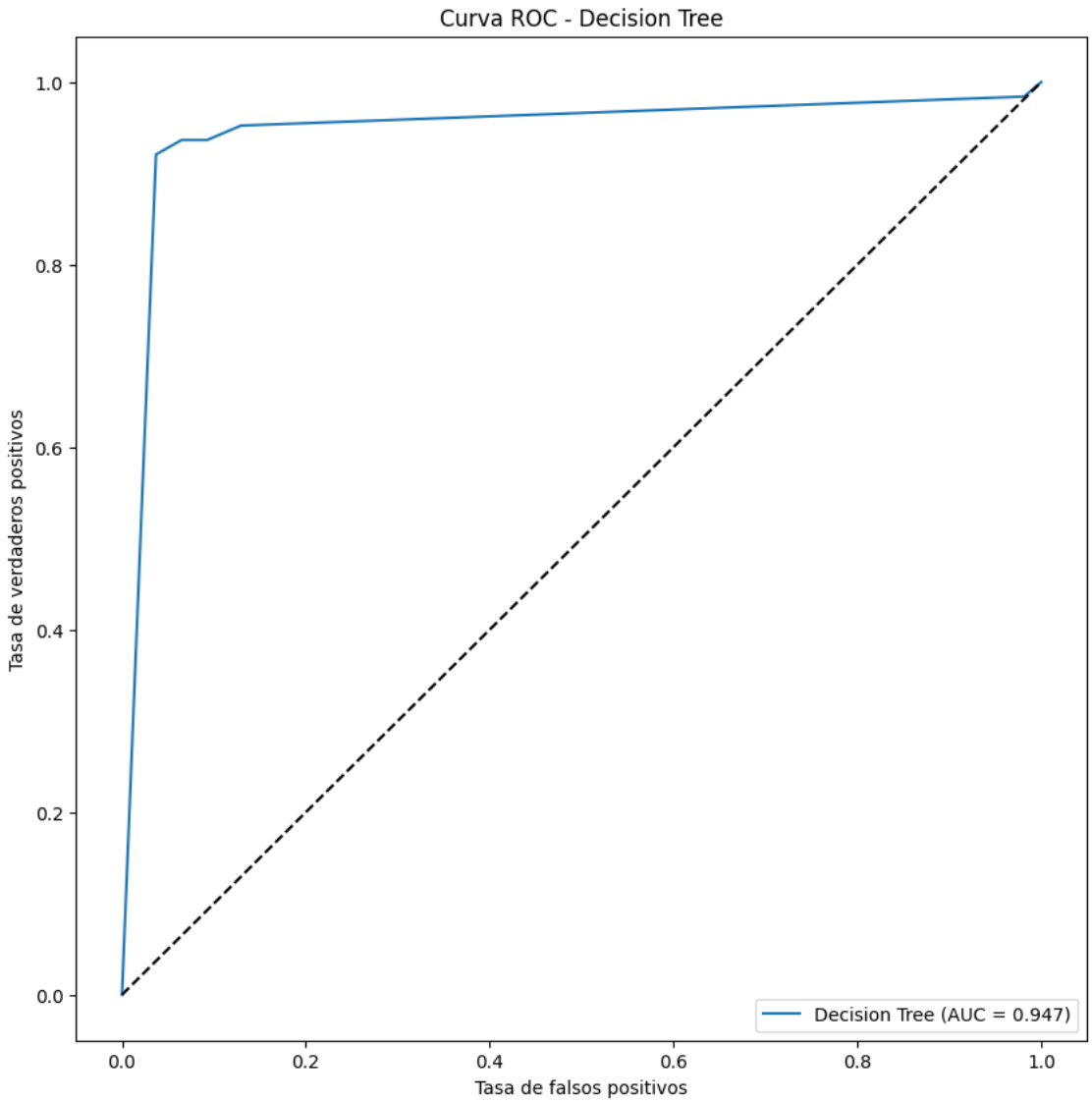
CLASIFICADOR NEAREST NEIGHBORS

El modelo KNN fue optimizado utilizando 'GridSearchCV' para encontrar el mejor número de vecinos. La mejor exactitud alcanzada fue de 0.9683 mediante 'n_neighbors = 7'.

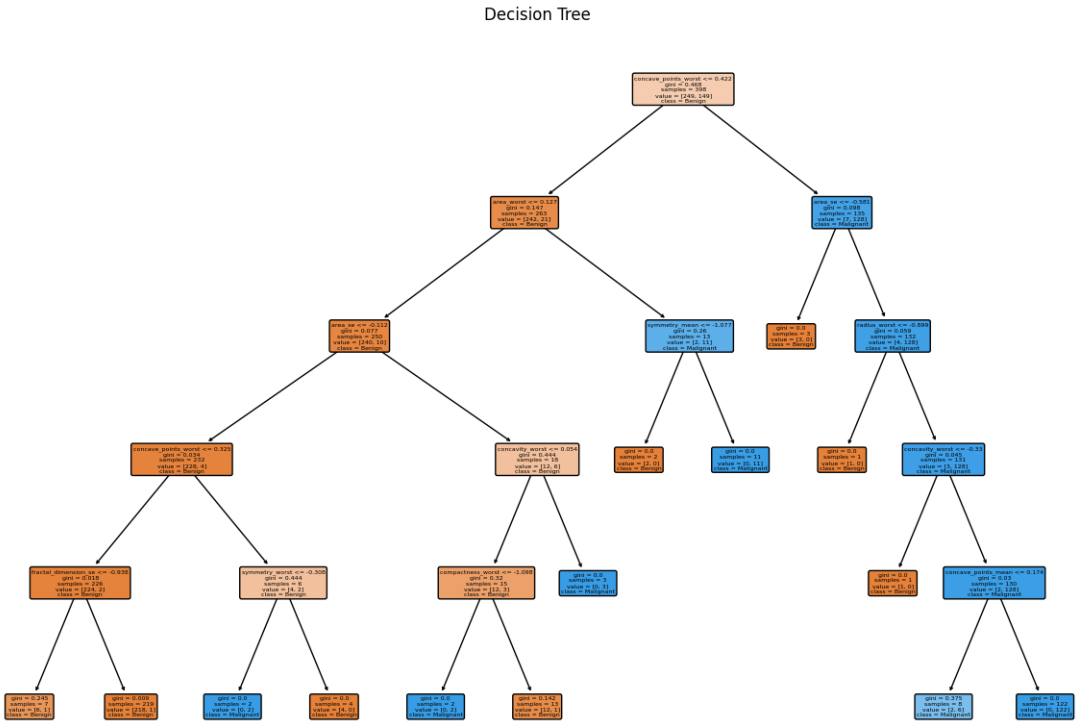


CLASIFICADOR DE ÁRBOL DE DECISIÓN

El Árbol de Decisión fue optimizado utilizando 'GridSearchCV' para los parámetros 'max_depth' (profundidad máxima del árbol) y 'min_samples_split' (mínimo de número de muestras por hoja). La mejor exactitud fue de 0.9227 con 'max_depth = 5' y 'min_samples_split = 5'.

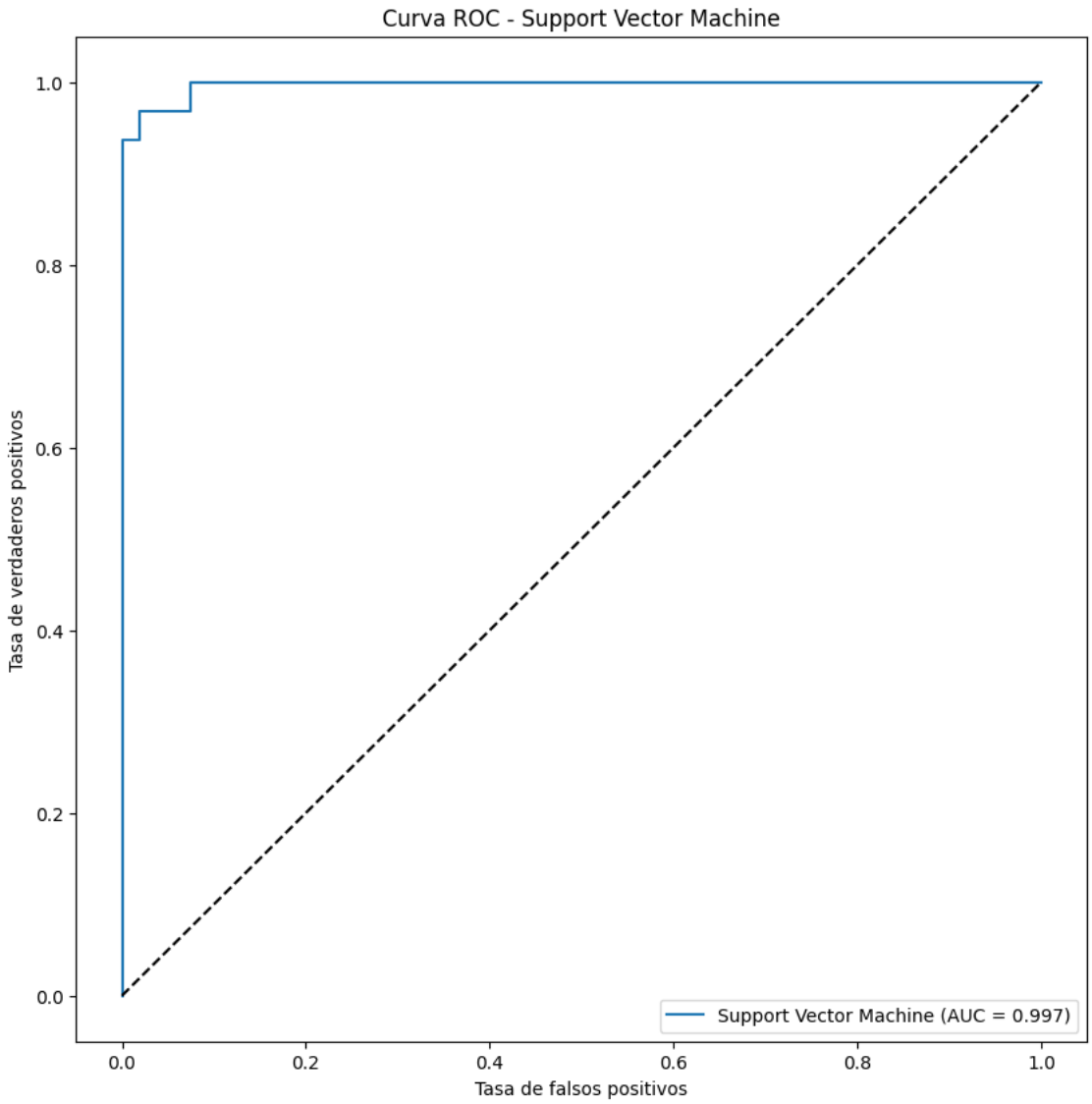


Como resultados obtenemos:



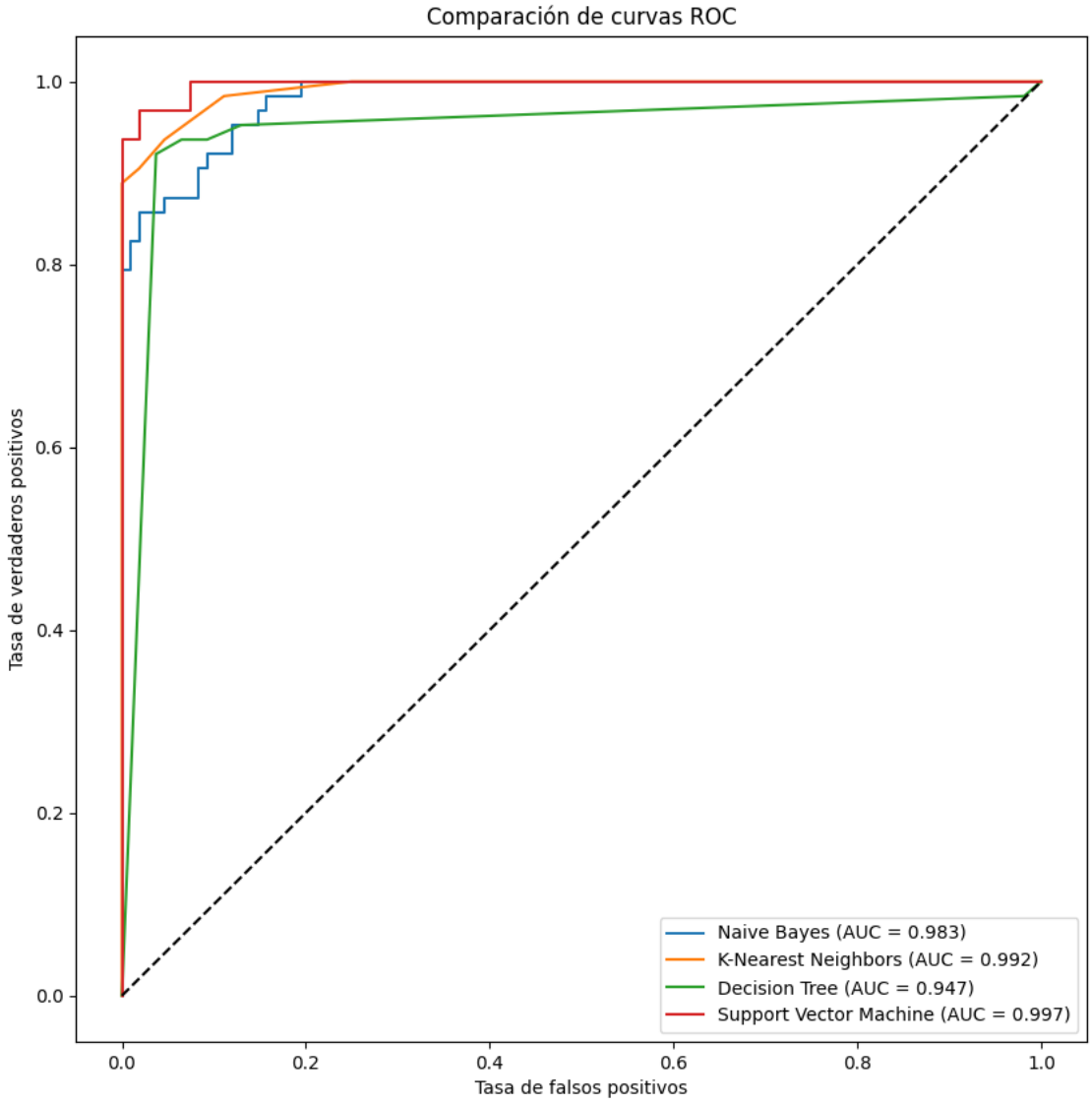
CLASIFICADOR SUPPORT VECTOR MACHINE

El modelo SVM fue optimizado utilizando ‘GridSearchCV’ para los parámetros ‘kernel’, ‘C’ y ‘gamma’. La mejor exactitud alcanzada fue de 0.9788 con ‘kernel = linear', ‘C = 0.1’, y ‘gamma = scale'.


Curva ROC - Support Vector Machine
Tasa de verdaderos positivos
Tasa de falsos positivos
0.0
0.2
0.4
0.6
0.8
1.0
Support Vector Machine (AUC = 0.997)

COMPARACIÓN CURVAS DE ROC

Ahora se presentan las curvas ROC para los clasificadores evaluados:



EVALUACIÓN FINAL

Los resultados indican que los modelos evaluados tienen un rendimiento alto según exactitud y AUC:

MODELO	EXACTITUD	AUC
Naive Bayes	0.9315	0.983
Nearest Neighbors	0.9683	0.992
Árbol De Decisión	0.9227	0.947
Support Vector Machine	0.9788	0.997

Esto sugiere que los modelos **K-Nearest Neighbors** y **SVM** son los más adecuados para este conjunto de datos específicos. Sin embargo, el modelo **Support Vector Machine (SVM)** mostró el mejor rendimiento global con una exactitud de 0.9758 y un AUC de 0.997.

REFERENCIAS

Enlace a Colab.

Enlace extra a mi primer proyecto personal sobre IA.

Repositorio.