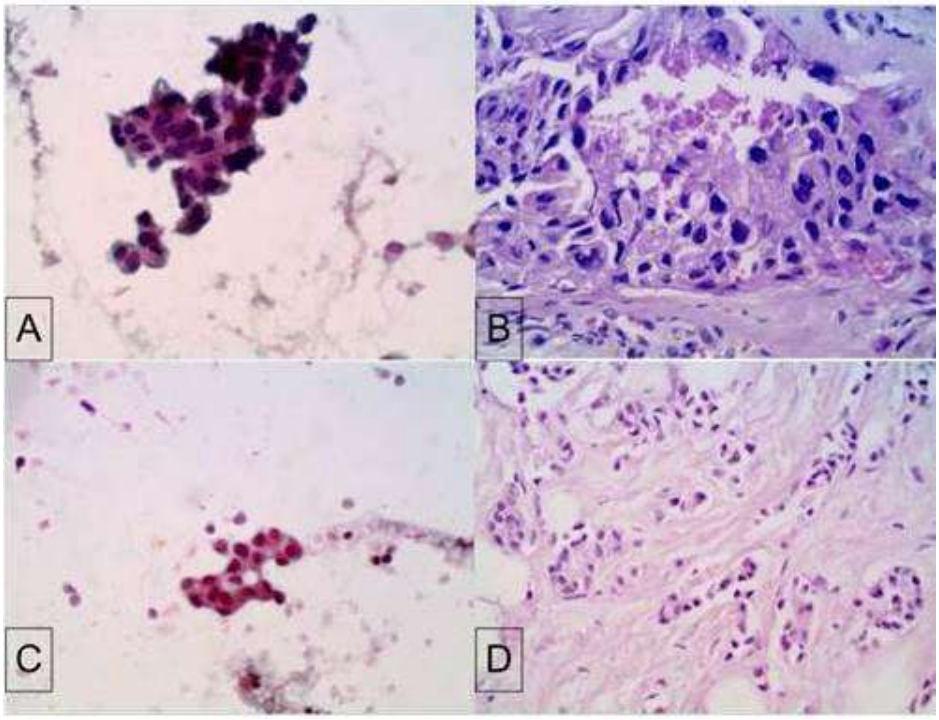
Clasificación de Cáncer de Mama usando KNN

Descripción del problema:

El objetivo es clasificar los tumores mamarios como malignos o benignos utilizando un modelo K-Nearest Neighbors (KNN) basado en las características de los núcleos celulares.

Fuente: Manuel Romera. Cáncer de Mama - Infografías Enlace



Fuente: Andrés Duque, Ana Karina Ramírez, Jorge Pérez. Punción aspiración con aguja fina guiada por ultrasonido de nódulos mamarios de alta sospecha Enlace

Punción Aspiración por Aguja Fina (PAAF):

La Punción Aspiración por Aguja Fina (PAAF) es un procedimiento diagnóstico utilizado en la evaluación de lesiones mamarias. Consiste en la obtención de una muestra de células del tejido mamario mediante una aguja fina para su posterior análisis bajo el microscopio. La PAAF es mínimamente invasiva y se usa para distinguir entre tumores benignos y malignos.

Aplicaciones en el diagnóstico de cáncer de mama:

- **Detección temprana:** La PAAF ayuda en la detección temprana del cáncer de mama, permitiendo el diagnóstico y tratamiento oportunos.
- **Minimización de riesgos:** Ofrece una alternativa menos invasiva en comparación con biopsias más extensas, reduciendo el riesgo para la paciente.
- **Guía para el tratamiento:** La información obtenida a través de PAAF puede ayudar a los médicos a planificar el tratamiento más adecuado basado en el tipo y grado del cáncer.

Dataset:

Este dataset describe las características de los núcleos celulares presentes en una imagen obtenida a partir de una aspiración con aguja fina (AAF) de una masa mamaria.

- **Fuente Kaggle**: Wisconsin Breast Cancer Dataset
- **Fuente UC Irvine Machine Learning Repository**: Wisconsin Breast Cancer Diagnostic Dataset

Descripción del Dataset:

- **Número de registros**: 569
- **Nombres de las columnas y su descripción**:
  1. **id**: Número de identificación (no utilizado en el modelo).
  2. **diagnosis**: Diagnóstico (M = maligno, B = benigno).
  3. **radius_mean**: Media de distancias desde el centro hasta los puntos en el perímetro.
  4. **texture_mean**: Desviación estándar de los valores de escala de grises.
  5. **perimeter_mean**: Perímetro.
  6. **area_mean**: Área.
  7. **smoothness_mean**: Variación local en la longitud del radio.
  8. **compactness_mean**: (perímetro^2 / área - 1.0).
  9. **concavity_mean**: Severidad de las porciones cóncavas del contorno.
  10. **concave_points_mean**: Número de porciones cóncavas del contorno.
  11. **symmetry_mean**: Simetría.
  12. **fractal_dimension_mean**: "Aproximación de la línea de costa" - 1.
  13. **radius_se**: Desviación estándar del radio.
  14. **texture_se**: Desviación estándar de la textura.
  15. **perimeter_se**: Desviación estándar del perímetro.
  16. **area_se**: Desviación estándar del área.
  17. **smoothness_se**: Desviación estándar de la suavidad.
  18. **compactness_se**: Desviación estándar de la compacidad.
  19. **concavity_se**: Desviación estándar de la concavidad.
  20. **concave_points_se**: Desviación estándar de los puntos cóncavos.
  21. **symmetry_se**: Desviación estándar de la simetría.
  22. **fractal_dimension_se**: Desviación estándar de la dimensión fractal.
  23. **radius_worst**: Peor valor del radio.
  24. **texture_worst**: Peor valor de la textura.
  25. **perimeter_worst**: Peor valor del perímetro.
  26. **area_worst**: Peor valor del área.
  27. **smoothness_worst**: Peor valor de la suavidad.
  28. **compactness_worst**: Peor valor de la compacidad.
  29. **concavity_worst**: Peor valor de la concavidad.
  30. **concave_points_worst**: Peor valor de los puntos cóncavos.
  31. **symmetry_worst**: Peor valor de la simetría.
  32. **fractal_dimension_worst**: Peor valor de la dimensión fractal.

Tipos de Datos:

- **id**: Entero.
- **diagnosis**: Cadena de texto (M, B).
- **Otros atributos**: Valores reales (float).

Elección de k=3 en KNN:

El valor de k en KNN define el número de vecinos más cercanos que se consideran para determinar la clase de un punto de datos. La elección de k=3 es común porque proporciona un equilibrio entre suavizar el ruido en los datos y mantener un modelo sensible a la estructura subyacente. Con k=3, el modelo es menos propenso a sobreajustarse a un único vecino ruidoso, pero sigue siendo lo suficientemente sensible como para capturar la variabilidad en los datos.

Vista previa del dataset

	id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean
0	842,302	M	17.99	10.38	122.8	1,001	0.118
1	842,517	M	20.57	17.77	132.9	1,326	0.084
2	84,300,903	M	19.69	21.25	130	1,203	0.109
3	84,348,301	M	11.42	20.38	77.58	386.1	0.142
4	84,358,402	M	20.29	14.34	135.1	1,297	0.100

Desempeño del modelo: Árbol de Decisión

Exactitud: `0.9736842105263158`

Precisión: `0.9701298701298702`

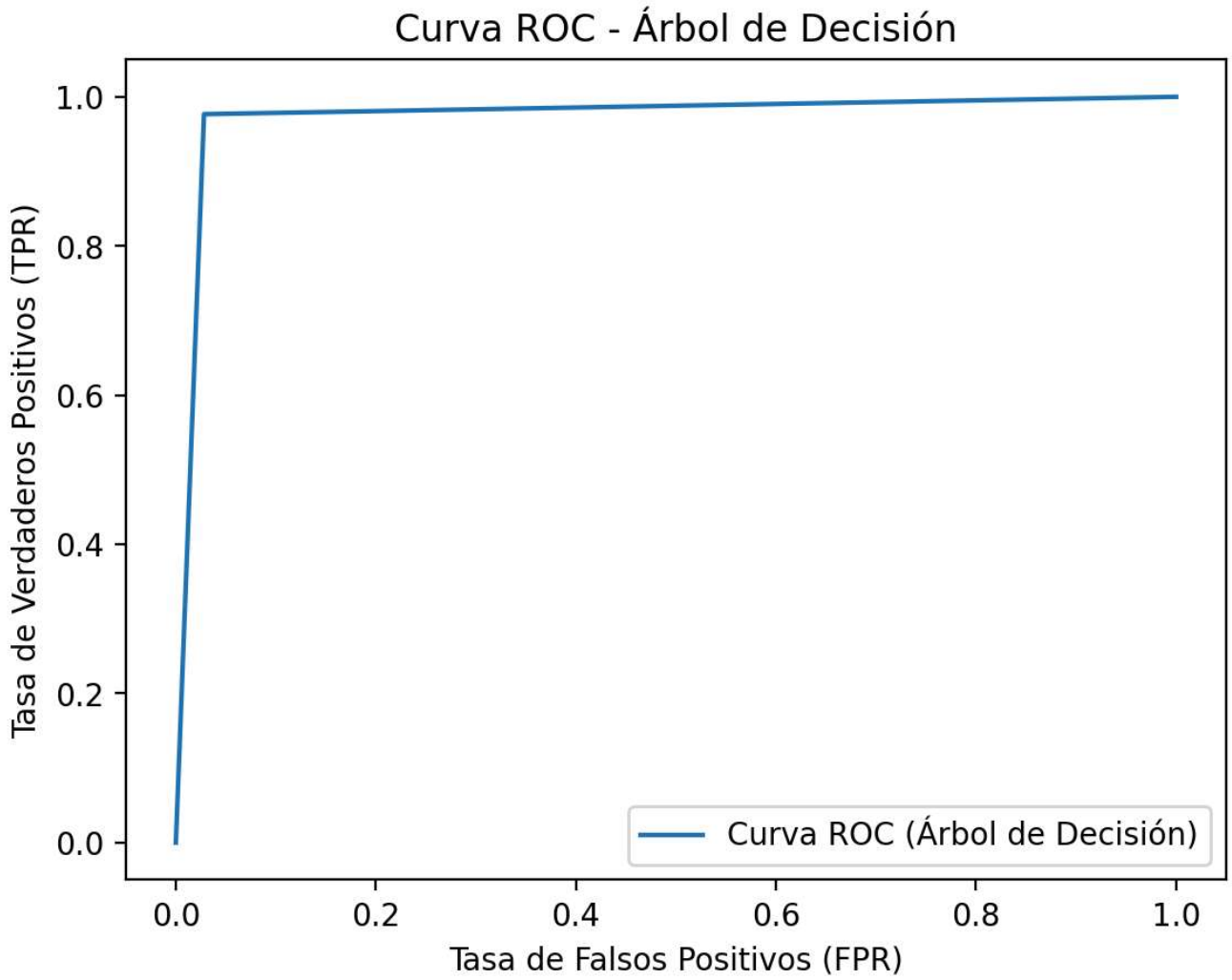
Recuperación (Recall): `0.9742875859810023`

F1 Score: `0.9721203228173148`

Matriz de Confusión:

0	1
69	2
1	42

Área Bajo la Curva (AUC): `0.9742875859810023`



Desempeño del modelo: Red Neuronal Multicapa (MLP)

Exactitud: `0.9736842105263158`

Precisión: `0.9742063492063492`

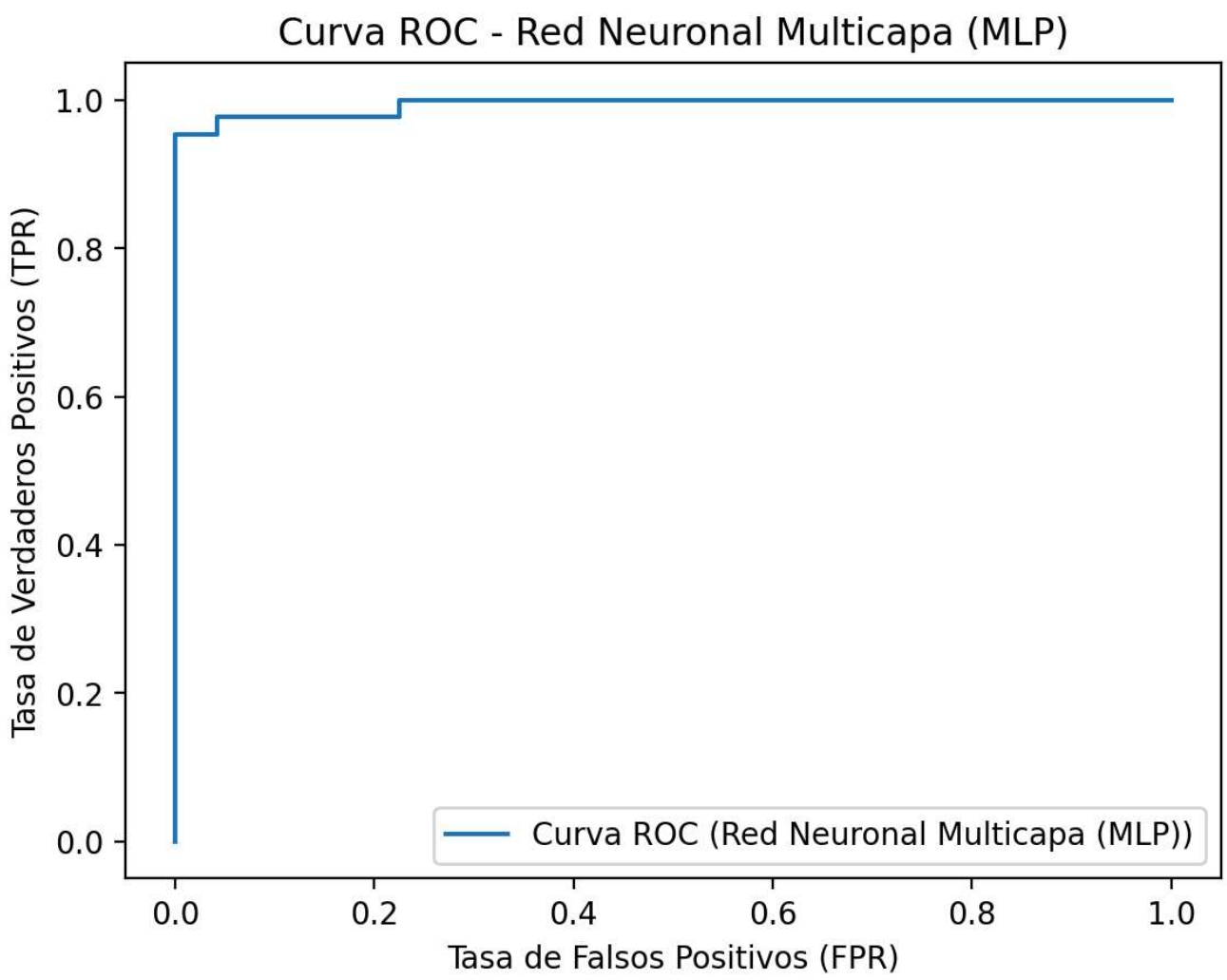
Recuperación (Recall): `0.9697019325253848`

F1 Score: `0.9718634306869601`

Matriz de Confusión:

0	1
70	1
2	41

Área Bajo la Curva (AUC): `0.9937766131673763`



Desempeño del modelo: Naive Bayes

Exactitud: `0.9649122807017544`

Precisión: `0.9672569328433009`

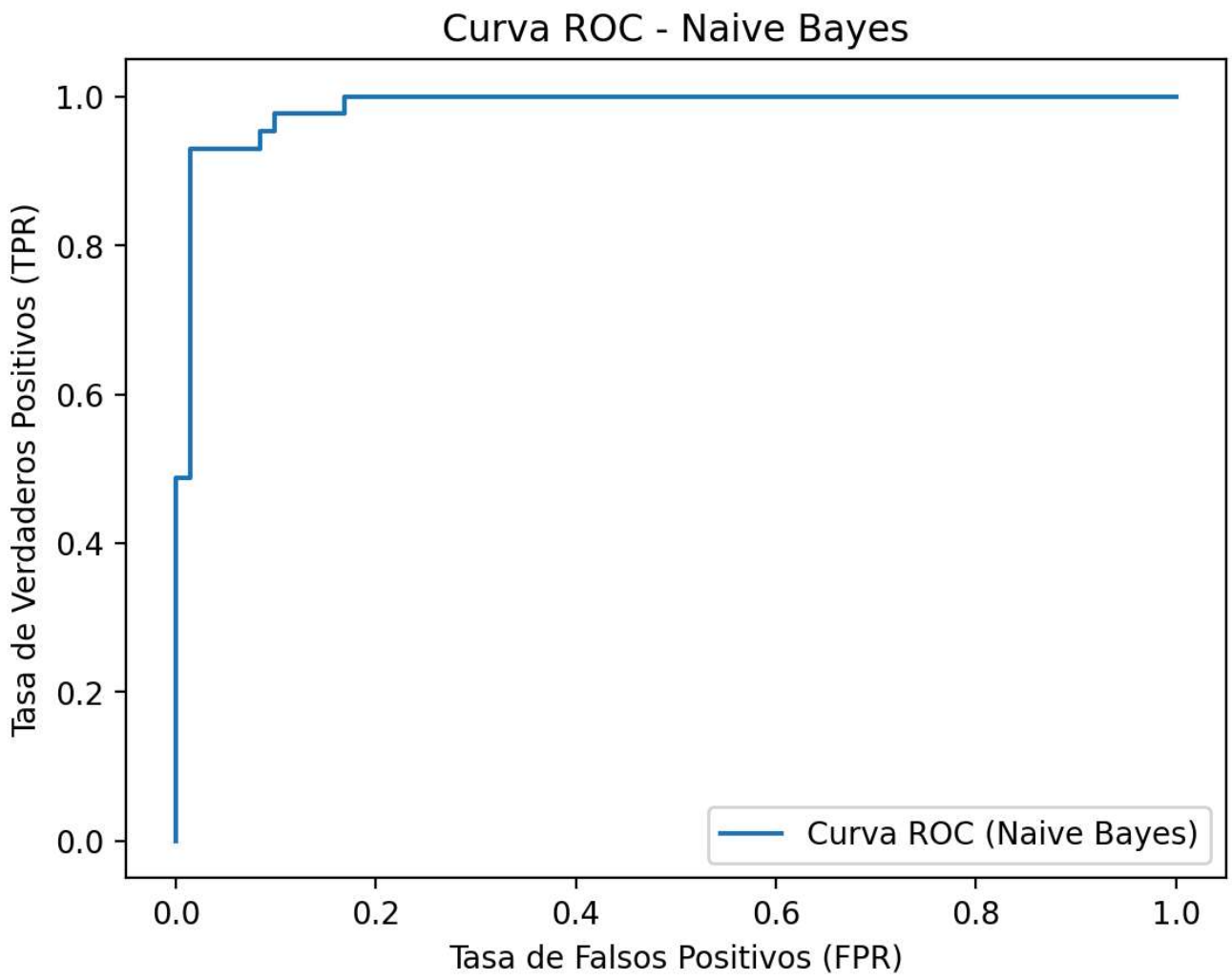
Recuperación (Recall): `0.9580740255486406`

F1 Score: `0.9623015873015872`

Matriz de Confusión:

0	1
70	1
3	40

Área Bajo la Curva (AUC): `0.9855879462823453`



Desempeño del modelo: K-Nearest Neighbors (KNN)

Exactitud: `0.956140350877193`

Precisión: `0.9516233766233766`

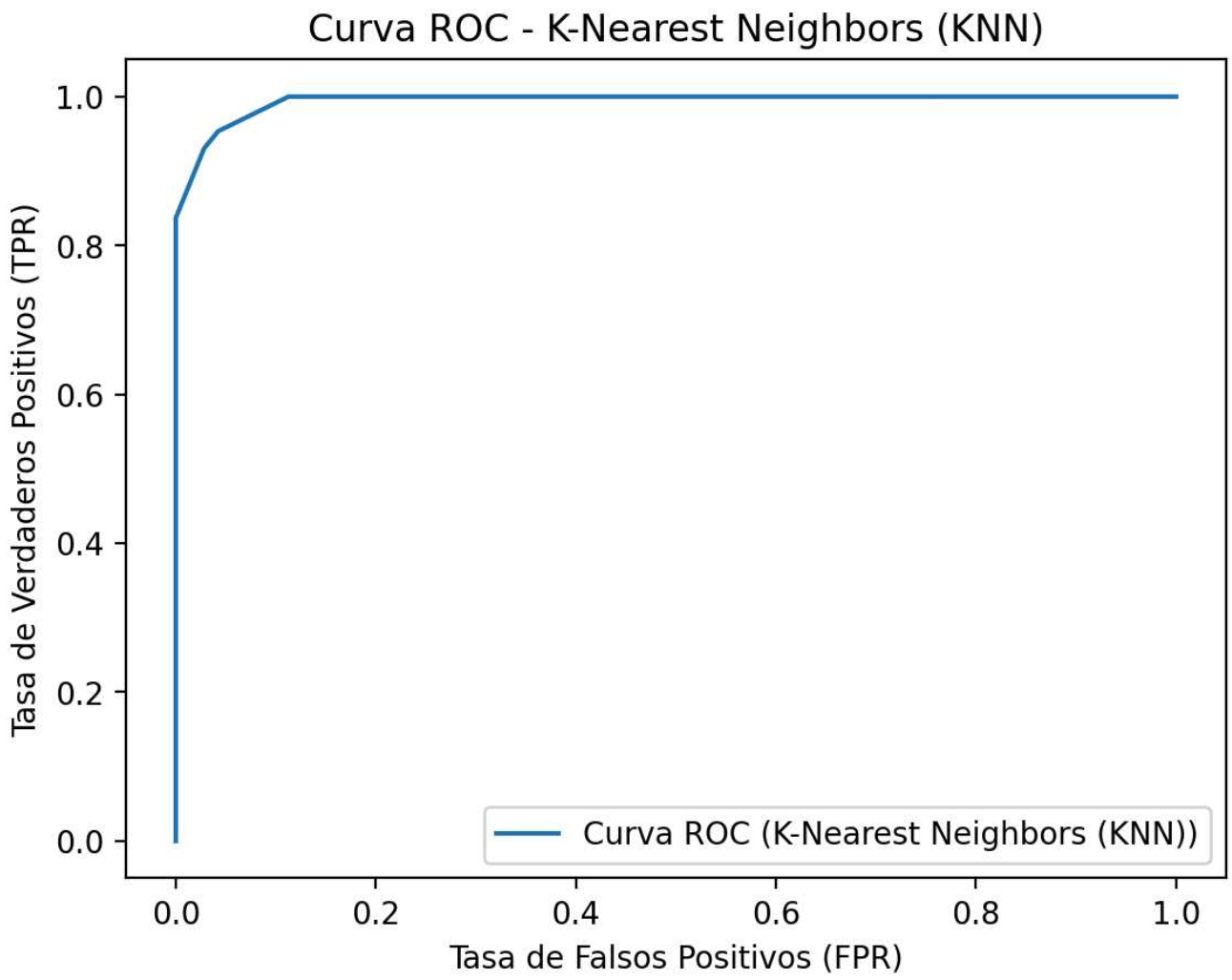
Recuperación (Recall): `0.9556174254831313`

F1 Score: `0.9535338713621913`

Matriz de Confusión:

0	1
68	3
2	41

Área Bajo la Curva (AUC): `0.9942679331804782`



Desempeño del modelo: Máquina de Vectores de Soporte (SVM - Lineal)

Exactitud: `0.9824561403508771`

Precisión: `0.981329839502129`

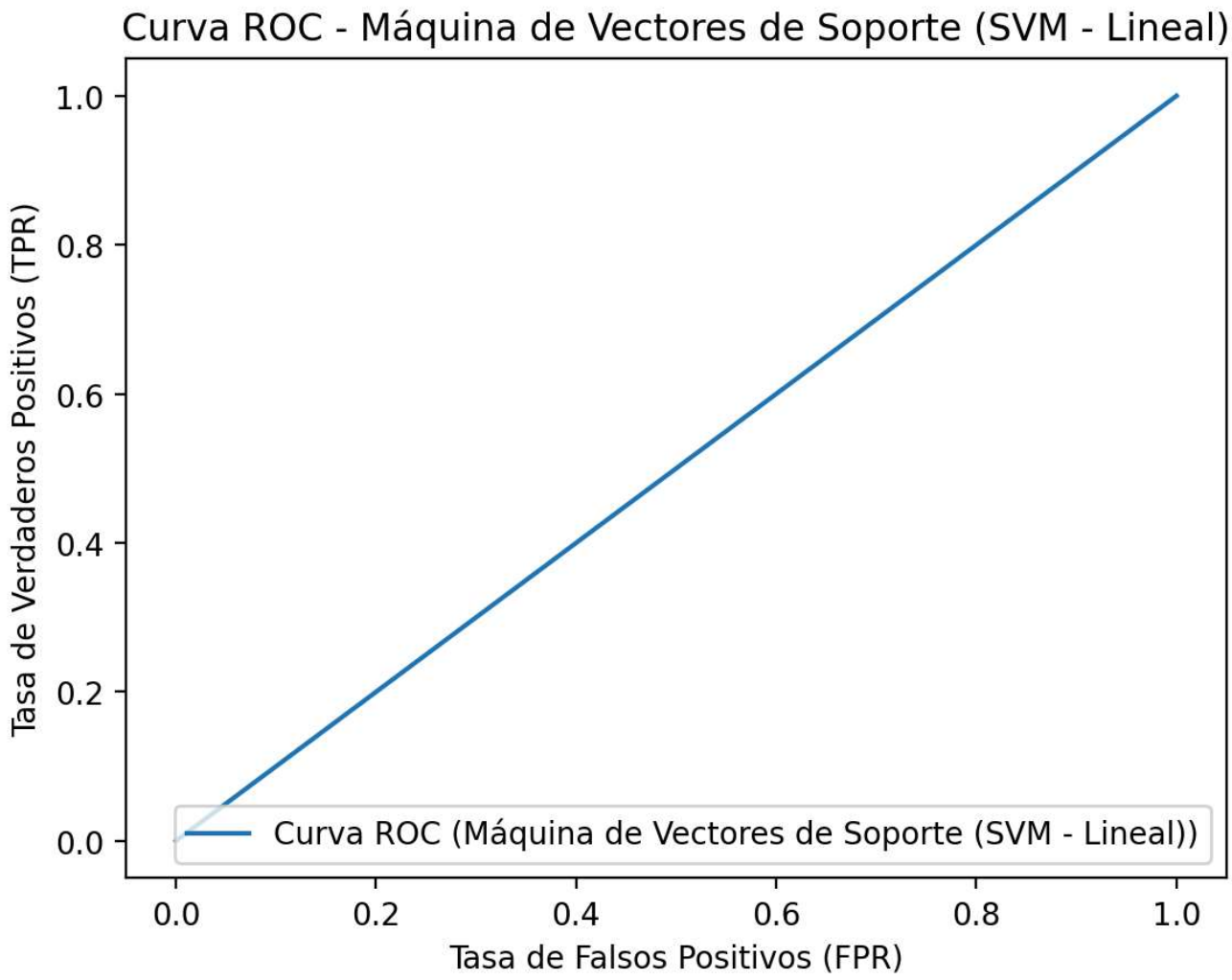
Recuperación (Recall): `0.981329839502129`

F1 Score: `0.981329839502129`

Matriz de Confusión:

0	1
70	1
1	42

Área Bajo la Curva (AUC): `0.5`



Desempeño del modelo: Máquina de Vectores de Soporte (SVM - RBF)

Exactitud: `0.9649122807017544`

Precisión: `0.9626596790042581`

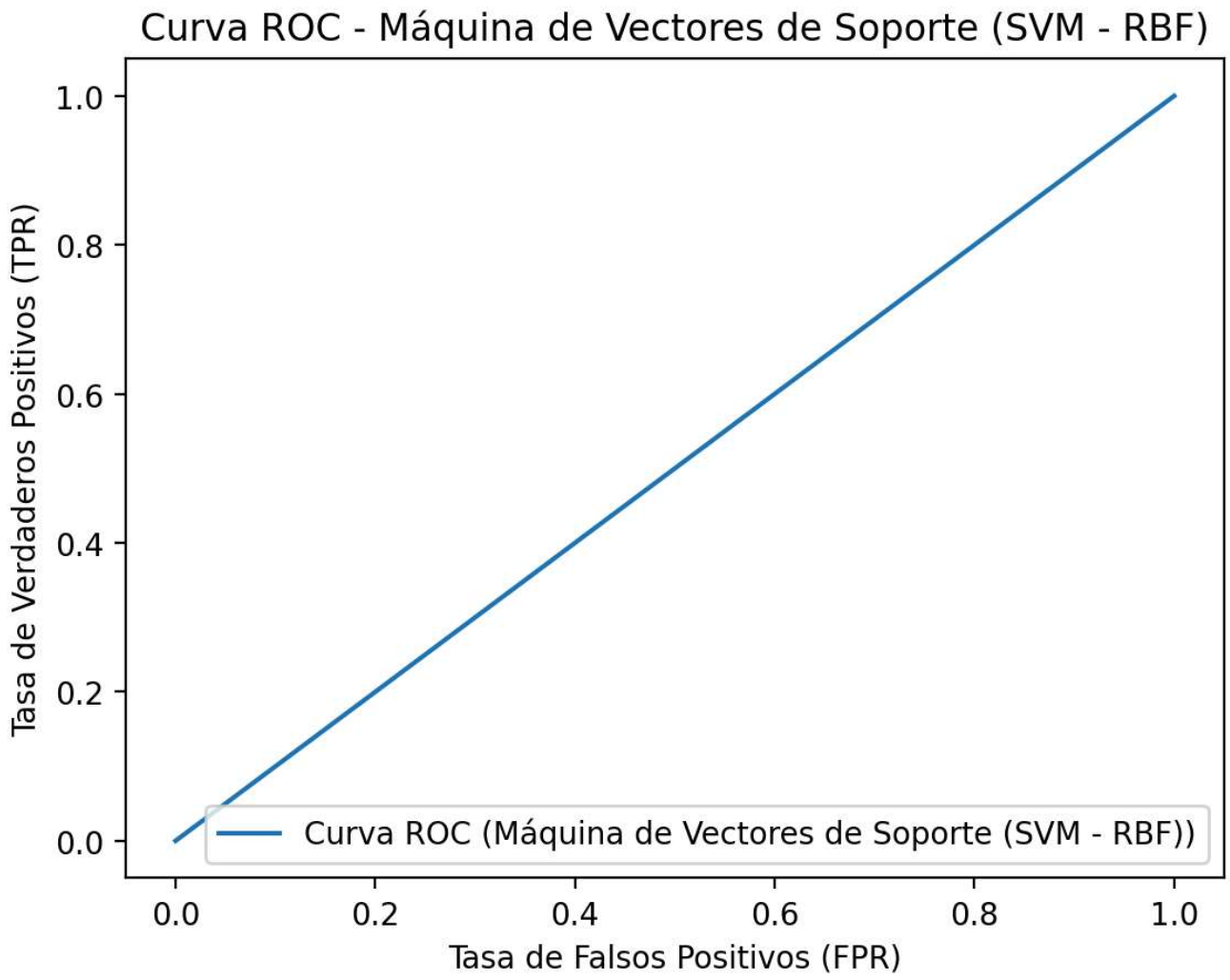
Recuperación (Recall): `0.9626596790042581`

F1 Score: `0.9626596790042581`

Matriz de Confusión:

0	1
69	2
2	41

Área Bajo la Curva (AUC): `0.5`



Desempeño del modelo: Random Forest

Exactitud: `0.9649122807017544`

Precisión: `0.9626596790042581`

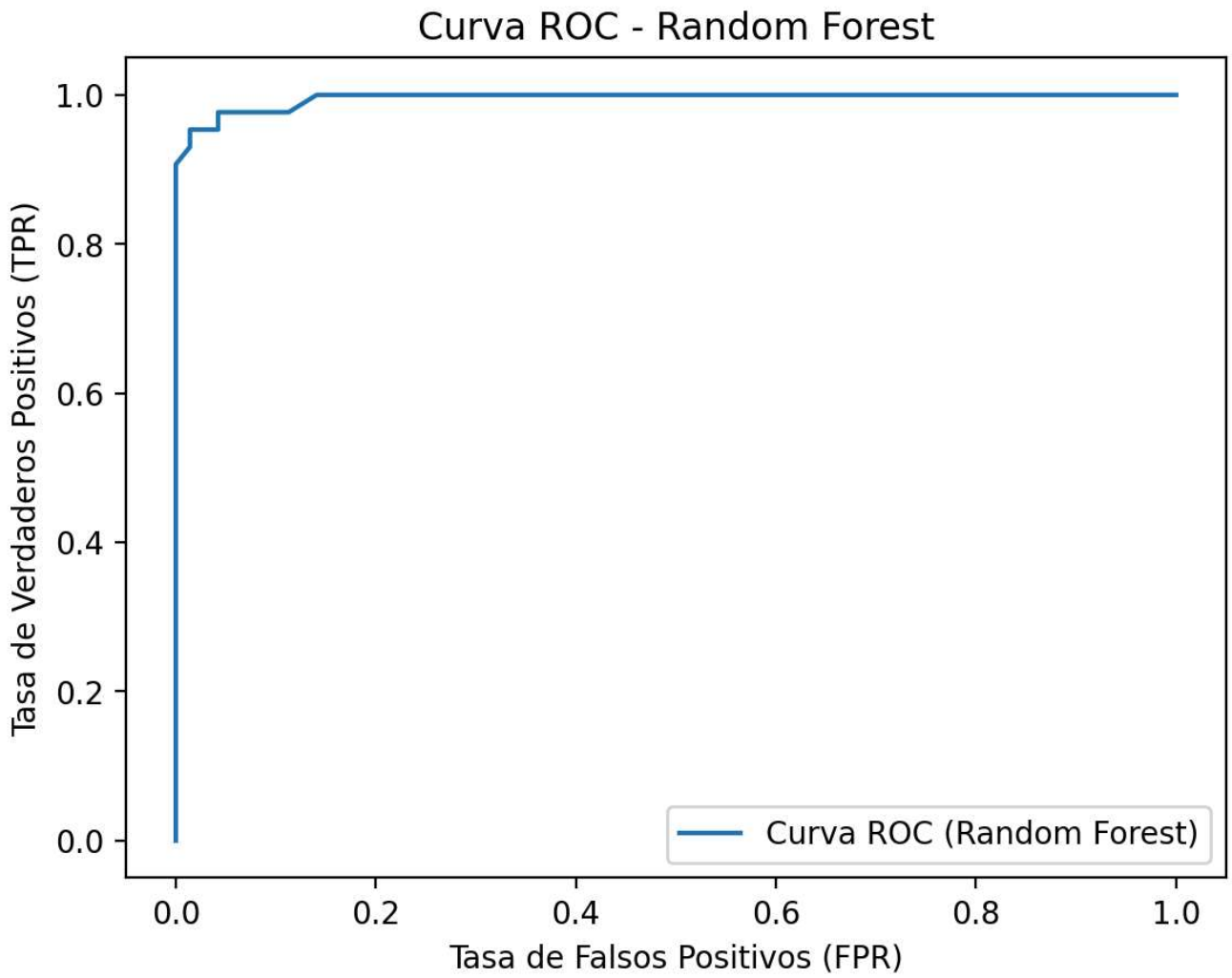
Recuperación (Recall): `0.9626596790042581`

F1 Score: `0.9626596790042581`

Matriz de Confusión:

0	1
69	2
2	41

Área Bajo la Curva (AUC): `0.9955781198820832`



Comparativa de Indicadores de Desempeño

	Modelo	Exactitud	Precisión	Recuperación	F1 Score	AUC
0	Árbol de Decisión	0.9737	0.9701	0.9743	0.9721	0.9743
1	Red Neuronal Multicapa (MLP)	0.9737	0.9742	0.9697	0.9719	0.9938
2	Naive Bayes	0.9649	0.9673	0.9581	0.9623	0.9856
3	K-Nearest Neighbors (KNN)	0.9561	0.9516	0.9556	0.9535	0.9943
4	Máquina de Vectores de Soporte (SVM - Lineal)	0.9825	0.9813	0.9813	0.9813	0.5
5	Máquina de Vectores de Soporte (SVM - RBF)	0.9649	0.9627	0.9627	0.9627	0.5
6	Random Forest	0.9649	0.9627	0.9627	0.9627	0.9956