INSTITUTO POLITÉCNICO NACIONAL

ESCUELA SUPERIOR DE CÓMPUTO

Aprendizaje de máquina e inteligencia artificial

Jiménez Alcantar Daniel

Práctica 03. Clasificación con vecinos más cercanos

Integrantes:
Cariño Sánchez Said
Quezada Espínola Paulina Yaelle
Verde Soria Pamela Aline

México CDMX a 12 de octubre de 2025

INTRODUCCIÓN.

El cáncer de mama es la primera causa de incidencia de mortalidad por cáncer en las mujeres adultas en el mundo y en América Latina. En 2020, a nivel mundial se presentaron 2,255,321 casos nuevos mientras que ocurrieron 683,502 defunciones del mismo cáncer en personas mayores de 25 años y más.

Cercanos a el día mundial contra el cáncer de mama (19 de octubre), decidimos usar el dataset "Breast Cancer Winsconsin (Diagnostic)" el cual cuenta con 30 medidas morfológicas derivadas de imágenes de núcleos celulares con los cuales es posible entrenar modelos de aprendizaje automático que discriminen entre tumores **malignos** y **benignos**. A pesar de ser un dataset donado en los 90's y que está situado en EUA, nosotros como equipo pudimos observar que posteriormente si quisiéramos podríamos realizar el mismo trabajo pero tomando un dataset actual y de México, pudiendo hacer uso incluso de uno de los modelos más sencillos de ML como lo es k-NN.

Siendo así esta práctica de gran ayuda para poder prácticar con datos reales y de gran utilidad en un ámbito de salud/social, esto con la finalidad de hacer conciencia sobre la importancia que tiene el diagnóstico oportuno cuando se presenta una enfermedad tan grave.

PROBLEMÁTICA.

El objetivo es maximizar el rendimiento global (accuracy, F1, ROC-AUC) sin perder de vista la sensibilidad (recall de malignos), que reduce el costo de FN.

En este trabajo se propone como línea base el clasificador de k vecinos más cercanos (k-NN) por su simplicidad, carácter no paramétrico y capacidad de capturar relaciones locales en el espacio de características.

Evaluamos variantes del algoritmo (voto estándar y ponderado por distancia), métricas de distancia, salidas probabilísticas y la selección sistemática de k mediante validación cruzada, con énfasis en controlar el riesgo de falsos negativos (FN) el cual sabemos es fundamental en un contexto clínico.

METODOLOGÍA.

Nuestra metodología consistió en:

- Carga y limpieza de datos; no se encontraron valores relevantes para consideración.
- Exploración de datos (EDA); estadísticas descriptivas donde principalmente se observa en la gráfica de pastel que la distribución de diagnósticos de tumores benignos es mayor que los malignos.
- Modelado; partición train/test para preservar la proporción de clases, preprocesamiento de StandardScaler dentro de un Pipeline para evitar fuga

de datos, búsqueda de hiperparámetros con GridSearchCV, evaluación en test (accuracy, ROC-AUC, etc.), análisis probabilístico para ajustar el umbral de decisión y visualización de los datos con matriz de confusión, curvas ROC y proyección PCA 2D.

MODELO ESTADÍSTICO.

Para un x nuevo, se buscan los k vecinos más cercanos por una distancia $d(\cdot,\cdot)$.

- Regla de decisión:
  Voto estándar (uniforme): $\hat{y} = \text{mode}(y_1, \dots, y_k)$
  Voto ponderado por distancia: pesos típicos $w_i = \frac{1}{d(x, x_1)+\varepsilon}$ (ó $\frac{1}{d^2}$), atiende más a vecinos muy cercanos.
  Salida probabilística: $\hat{P}(y = 1|x) = \frac{\sum_i w_i 1(y_1=1)}{\sum_i w_i}$ útil para umbralizar por costos.
- Métricas de distancia (espacio estandarizado):
  Euclídea (L2): sensible a diferencias grandes, suele ser el baseline.
  Manhattan (L1): más robusta a outliers, es útil si algunas features presentan colas pesadas.
  Chebyshev (L∞): penaliza la peor discrepancia, es útil si "una gran diferencia en una sola variable" basta para considerar "no similar".
- Sesgo-varianza y elección de k:
  k pequeño → baja frontera de decisión (alta varinza, posible sobreajuste).
  k grande → suaviza ruido (más sesgo, posible infraajuste).
  Elegir k por CV buscando el "codo" en k vs accuracy/F1.

MODELO COMPUTACIONAL.

El notebook implementa el modelo en Python utilizando librerías como; *pandas* y *numpy* para manipulación de datos, *matplotlib* ¿y *seaborn?* para visualización, *scikit-learn* para modelado y evaluación.
El flujo se estructuró con un Pipeline que integra StandardScaler (estandarización de variables) y KneighborsClassifier (k-NN).
La selección de hiperparámetros se realizó mediante GridSearchCV con validación cruzada estratificada (StratifiedKFold, 5 folds, random_state=42), explorando: número de vecinos $k \in \{1, 3, 5, 7, 9, 11, 15, 21, 31\}$,
esquema de voto (weights ∈ {uniform, distance}) y
métrica de distancia (Minkowski con $p \in \{1, 2, 3\}$ y Chebyshev).
El mejor modelo (según exactitud promedio en CV) se reajustó automáticamente sobre el conjunto d entrenamiento y se evaluó en prueba con accuracy, F1 (binario y macro), ROC-AUC y log-loss; adicionalmente se reportó la matriz de confusión (ConfusionMatrixDisplay) y las curvas ROC (RocCurveDisplay).

Las salidas probailísticas del clasificador (predict_proba) permiten, si es requerido, ajustar el umbral de decisión conforme a criterios clínicos (priorización de sensibilidad).
Para interpretación geométrica se utilizó PCA (2 componentes) exclusivamente con fines de visualización de fronteras y distribución de aciertos/errores.

PROPUESTA DE SOLUCIÓN PARA EL USO DE CLASIFICACIÓN CON VECINOS MÁS CERCANOS PARA EVALUAR SU DESEMPEÑO CON:

1. *ESTÁNDAR:*
   (weights= 'uniform') es la línea base; evalúa el accuracy, F1 y la matriz de confusión; es útil para medir la ganancia de otras variantes.
2. *PONDERADO:*
   (weights='distance') atenúa votos de vecinos lejanos; puede mejorar en zonas con solapamiento de clases o densidades desiguales.
3. *DISTANCIAS VARIADAS:*
   Comparación entre L2 (suele ir bien tras escalado) vs L1 (robusta) vs L∞ (criterio umbral)
4. *PROBABILÍSTICO:*
   Con *predict_proba* evaluamos log-loss y ROC-AUC, lo que nos permite mover el umbral para favorecer sensibilidad. Por ejemplo, sensibilidad a ≥0.95 si el costo de FN es alto.
5. *SELECCIÓN DE K:*
   Curva k vs accuracy/F1 (CV) para ubicar el "codo". Reportar el mejor k y la métrica correspondiente (media±std en CV) y luego reentrenar en train para evaluar en test.

VISUALIZACIÓN DE LA SOLUCIÓN.

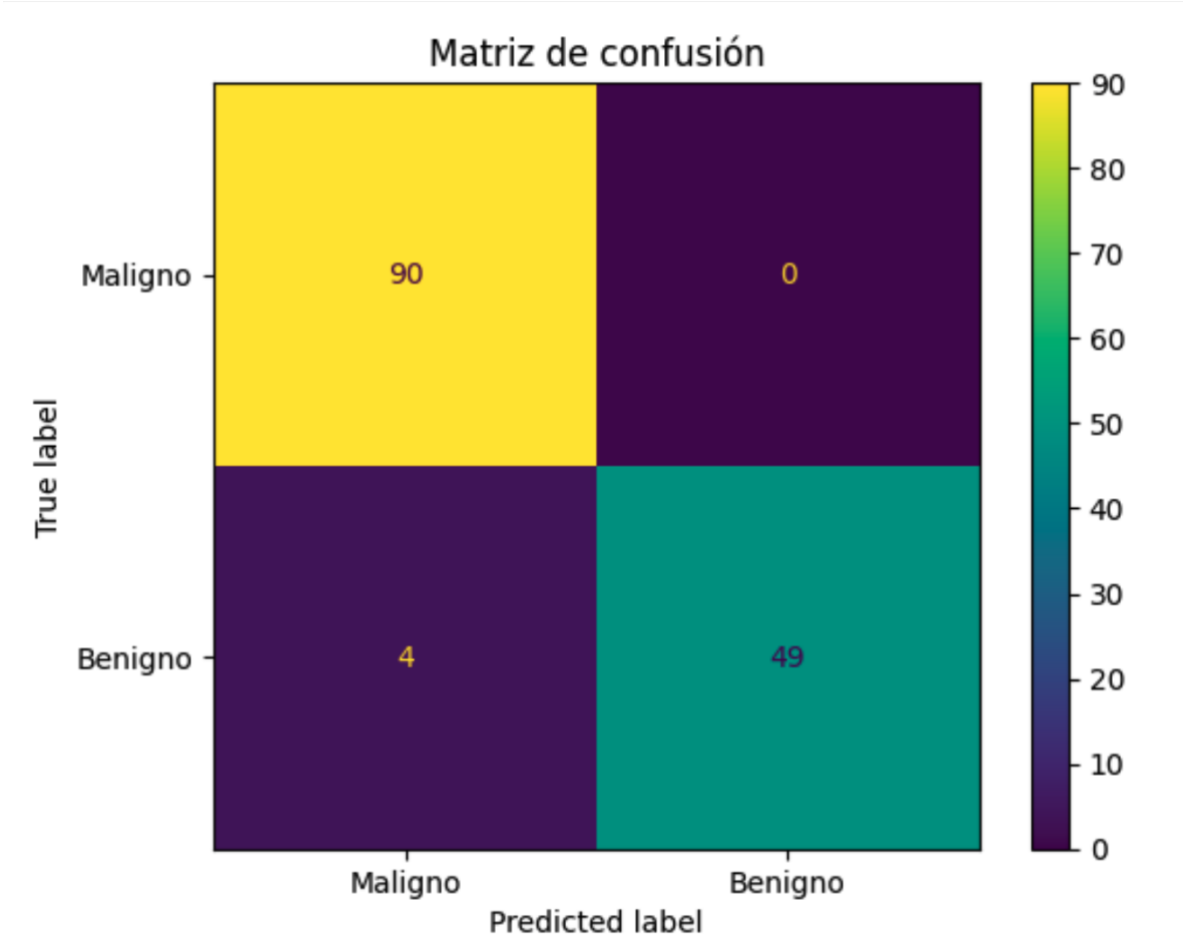
Matriz de confusión.

Nuestra matriz indica que el modelo:

Detecta todos los casos malignos (sin falsos negativos).

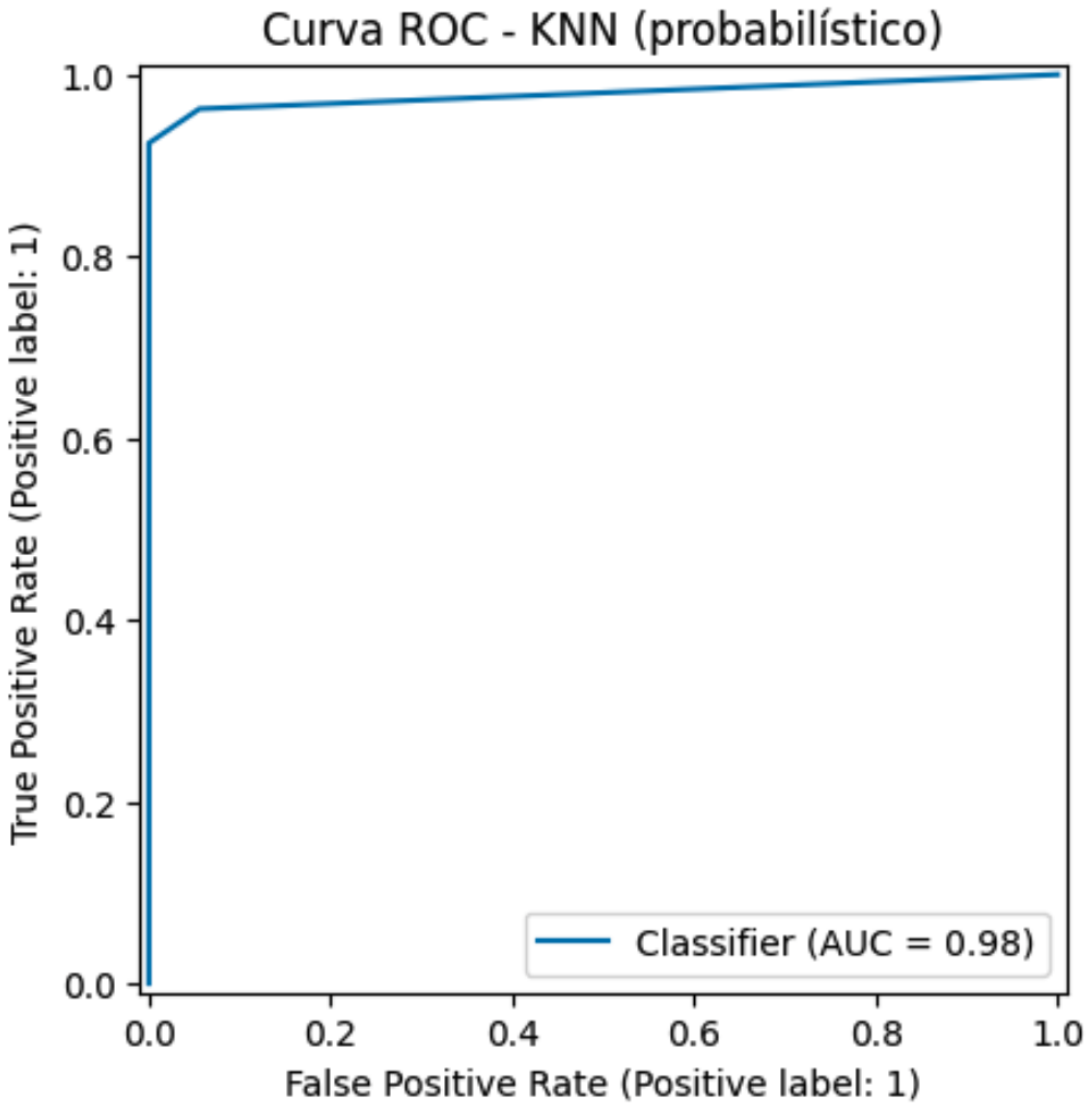
4 casos benignos fueron marcados como malignos (falsos positivos)

El modelo está sesgado a favor de la seguridad clínica porque no deja escapar malignos, a costa de pocos falsos positivos en benignos.



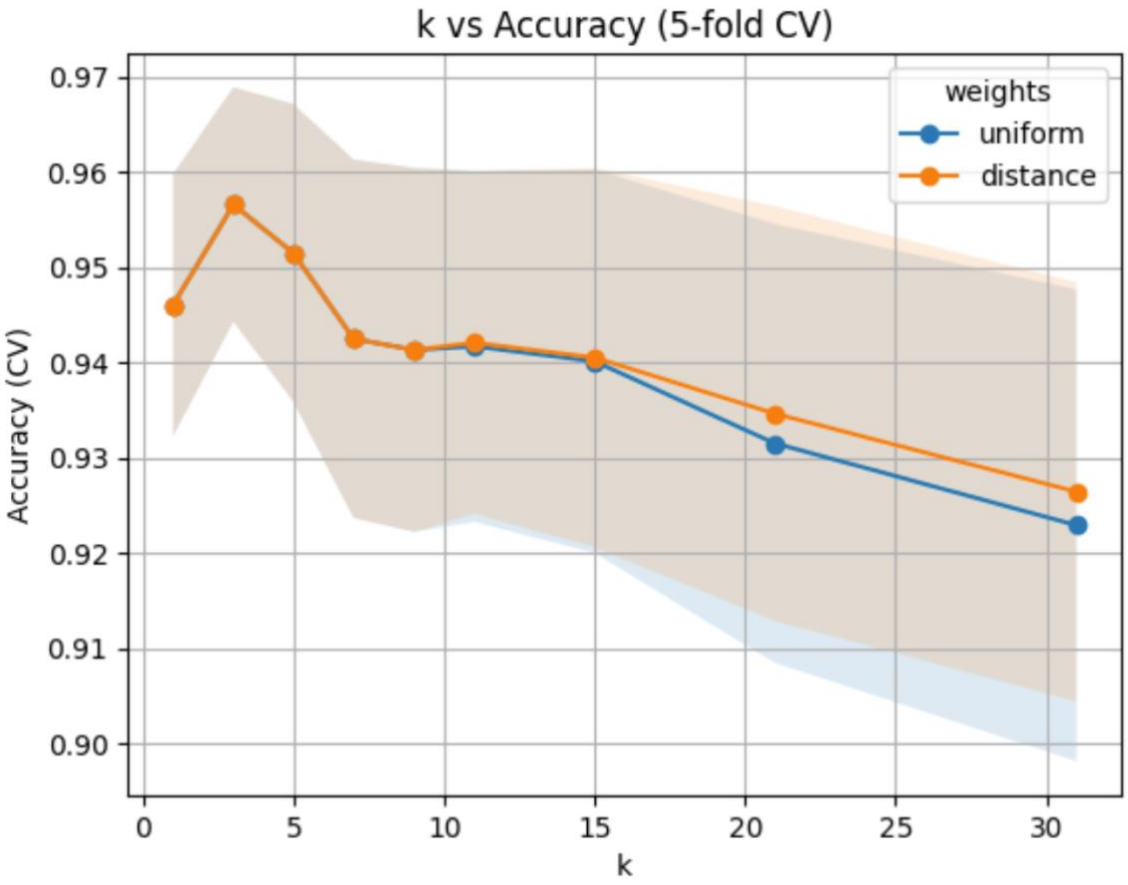
Curva de ROC.

La curva pegada al eje izquierdo y al techo indica alta capacidad de discriminación: el modelo separa muy bien malignos y benignos a lo largo de distintos umbrales. AUC= 98% nos inidca que, en un 98% de pares (maligno vs benigno al azar), el modelo asigna mayor probabilidad al caso maligno que al benigno.



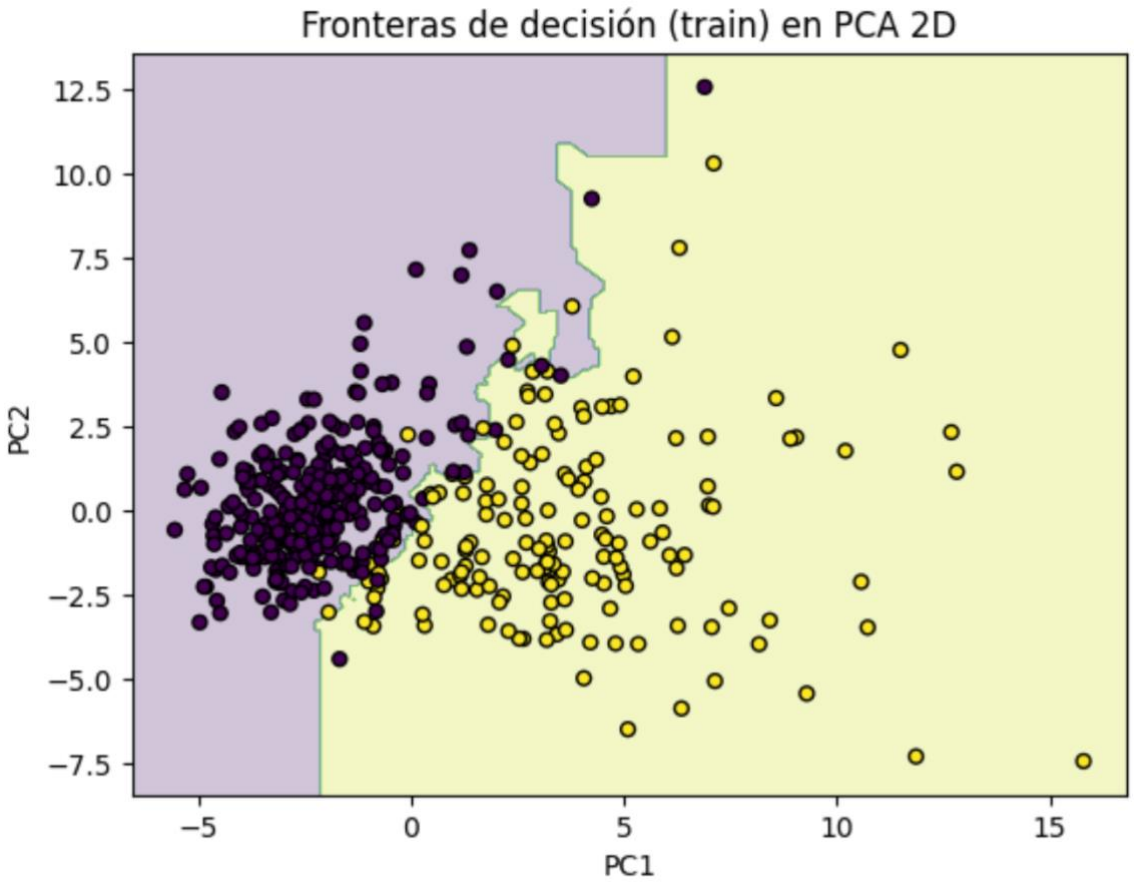
k vs accuracy.

El mejor rendimiento aparece con k pequeño (≈3-5), en especial con weights = distance, que supera levemente a uniform.
A partir de k≥ 7-10 la accuracy decrece de forma sostenida, entre más vecinos mayor sesgo/infraajuste tenemos.



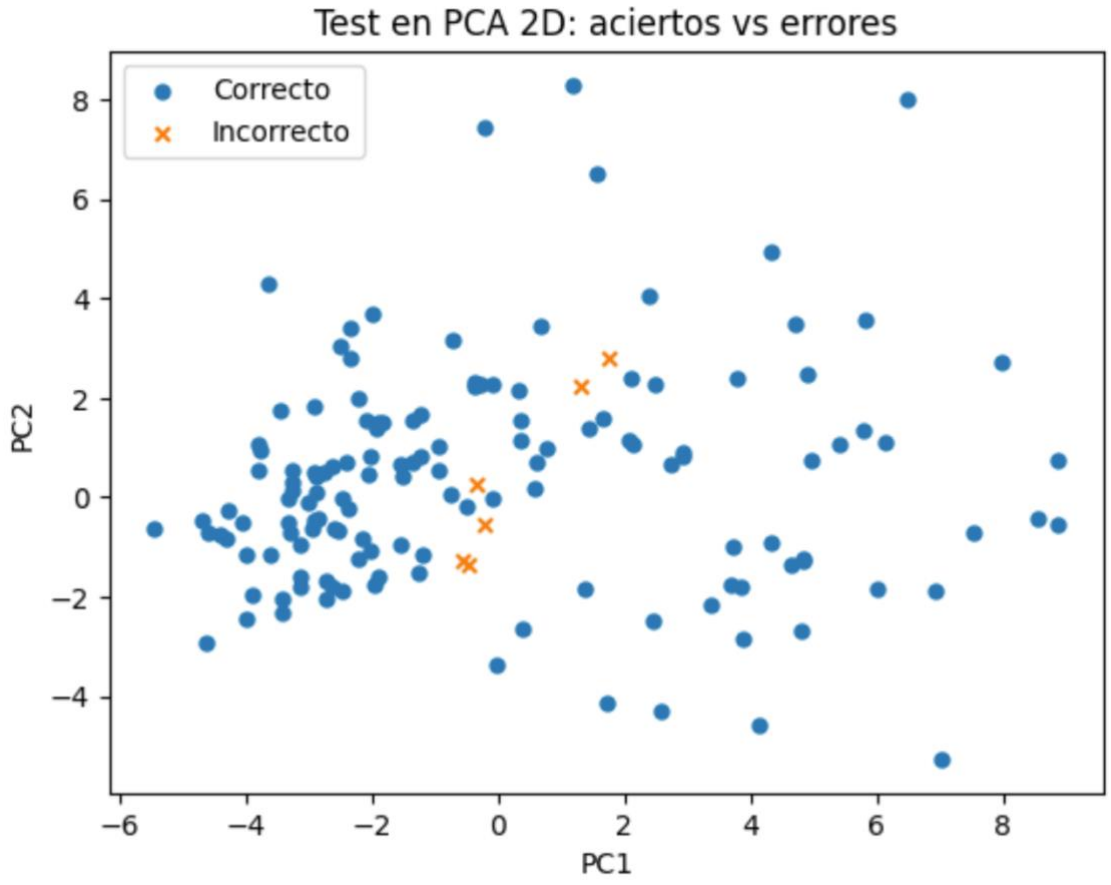
Fronteras de decisión.

Los puntos con PC1 alto concentran muestras con núcleos más grandes e irregulares, rasgos más compatibles con malignidad; los puntos de PC1 bajo agrupan morfologías más regulares y pequeñas, típicas de benignidad.
La franja de solapamiento cerca de la frontera corresponde a casos borderline (características intermedias) donde es esperable mayor incertidumbre diagnóstica. Clínicamente, esos son los casos que requieren corroboración como criterios adicionales.
Fuera de esa zona, la separación es amplia: la mayoría de las muestras presentan patrones morfológicos claros y coherentes con su clase clínica.



Aciertos vs Errores.

Los errores se concentran en una zona central del plano (PC1^PC2 ≈ 0-2) donde las muestras presentan patrones morfológicos intermedios y el caso es más ambiguo.
Lejos de esa franja, los puntos correctos dominan: hay buena separabilidad cuando las características proyectadas son claramente bajas/altas.
Nuevamente, en un ámbito clínico, los casos en la región central son borderline y serían candidatos a criterios adicionales para reducir equivocaciones en esa zona.



CONCLUSIONES.

Cariño Sánchez Said.

En esta práctica pudimos aplicar lo que hemos aprendido sobre clasificación con k-NN en un caso con impacto real, usando el dataset de cáncer de mama. Vimos cómo, a pesar de ser un modelo sencillo, puede ofrecer buenos resultados si se hace una buena limpieza, escalado y selección de hiperparámetros. Nos llamó la atención que el modelo logró detectar todos los casos malignos, lo cual es muy importante en un contexto clínico, aunque con algunos falsos positivos que son aceptables considerando el tipo de problema.

En general, fue una experiencia útil porque nos permitió conectar la parte técnica con algo más humano y tangible. Aprendimos a interpretar las métricas, a usar herramientas como GridSearchCV y a entender mejor cómo elegir el valor de k. Además, nos ayudó a ver que la ciencia de datos puede tener un impacto real en la salud y la toma de decisiones.

Quezada Espínola Paulina Yaëlle.

El clasificador k-NN constituye una línea base metodológicamente válida para el dataset Breast Cancer Winsconsin (Diagnostic). Tras la estandarización de variables y la búsqueda sistemática de hiperpárametros (número de vecinos, esquema de voto y métrica de distancia) mediante validación cruzada, se obtuvo un desempeño elevado y estable -reflejado en valores altos de exactitud y área bajo la curva ROC- acompañado de una tasa reducida de falsos negativos en las configuraciones óptimas. En términos prácticos, las distancias Euclídea y Manhattan (Minkowski) presentan un comportamiento consistentemente favorable después del escalado; por su parte, el voto ponderado por distancia tiende a prodcuir mejoras acotadas en regiones fronterizas. La selección del número de vecinos mediante el método del codo permite equilibrar adecuadamente el compromiso sesgo-varianza, evitando tanto el sobreajuste asociado a valores pequeños de k como el infraajuste derivado de k excesivamente grandes. Adicionalmente, el empleo de salidas probabilísticas facilita el ajuste de umbrales de decisión con el fin de priorizar sensibilidad clínica, manteniendo pérdidas controladas de especifidad.

El conjunto de trabajo Pipeline, GridSearchCV con métricas y visualizaciones (matriz de confusión, ROC, k vs rendimiento), PCA 2D) es riguroso y reproducible, sirve como base para comparar con modelos más complejos y para ajustar umbrales probabilísticos según criterios clínicos.

Verde Soria Aline Pamela.

Haber utilizado el algoritmo k-NN fue muy interesante en este caso, ya que permitió crear una herramienta computacional que puede ayudar a diagnosticar el cáncer de mama, una enfermedad que actualmente afecta a miles de mujeres y cuya detección oportuna puede salvar vidas. Al generar las métricas de evaluación obtuvimos un AUC bastante alto y sin falsos negativos, lo que refleja que este algoritmo realizó la clasificación de los casos de manera muy eficiente. Esto significa que la detección de casos malignos fue muy precisa, algo fundamental cuando se trata de vidas humanas. Aunque se observaron algunos falsos positivos, estos representan un riesgo mucho menor.

Además, el uso de la validación cruzada nos permitió seleccionar un número óptimo de vecinos (k) y entender cómo este parámetro influye en el equilibrio entre el sesgo y la varianza, evitando tanto el sobreajuste como el infraajuste.

En general, aunque suele pensarse que este algoritmo es simple, la práctica nos permitió confirmar que puede ser muy potente para realizar tareas de clasificación tan complejas como los diagnósticos médicos.