UNIVERSIDAD TECNOLÓGICA DE CHIHUAHUA

Tecnologías de la información

Extracción de Conocimiento en Bases de Datos

Reporte de Métricas de Evaluación

IDGS91N

PROFESOR:
Enrique Mascote

Alumno:
Emanuel Chavira

29 de Noviembre de 2025

III.2. Reporte de Métricas de Evaluación

1. Introducción

El desarrollo de modelos de aprendizaje automático requiere evaluar su desempeño con métricas apropiadas. Existen métricas específicas para tareas de clasificación (cuando se predicen categorías) y para regresión (cuando se predicen valores numéricos continuos). Este reporte tiene como objetivo investigar diversas métricas de evaluación para modelos de clasificación y regresión, y aplicar un clasificador K-Nearest Neighbors (KNN) a un conjunto de datos con variables de glucosa y edad para predecir una etiqueta binaria. Se presentan los pasos de preprocesamiento, entrenamiento y evaluación del modelo, así como un análisis de los resultados obtenidos.

2. Investigación de métricas

2.1 Métricas de clasificación

- **Exactitud (accuracy)**: mide la proporción de clasificaciones correctas sobre el total de ejemplos. Matemáticamente se define como
  (Accuracy = (TP + TN) / (TP + TN + FP + FN)), donde *TP*, *TN*, *FP* y *FN* son los verdaderos positivos, verdaderos negativos, falsos positivos y falsos negativos, respectivamente.
  Interpretación: indica qué fracción de las predicciones totales coincide con la realidad. Su uso es adecuado en conjuntos equilibrados, pero puede ser engañoso cuando una clase es mayoritaria.
  Ventajas: fácil de entender y de comunicar; resume el rendimiento general del modelo.
  Limitaciones: en datos desequilibrados puede ofrecer una visión optimista porque ignora la distribución de clases; no distingue entre tipos de error.

- **Precisión (precision)**: cuantifica la fracción de predicciones positivas que son realmente positivas. Se define como
  (Precision = TP / (TP + FP)).

Interpretación: responde a la pregunta "¿de todas las veces que el modelo predijo positivo, cuántas acertó?". Es útil cuando el coste de un falso positivo es alto.

Ventajas: penaliza los falsos positivos y por ello es relevante en aplicaciones donde se desea minimizar falsas alarmas.

Limitaciones: puede ser elevada aun cuando el modelo no identifique todos los positivos; no considera los falsos negativos.

- **Recuperación (recall o sensibilidad)**: mide la proporción de positivos reales que se clasifican correctamente. Su fórmula es
  (Recall = TP / (TP + FN)).
  Interpretación: indica qué tan bien el modelo detecta todos los casos positivos. Es prioritaria cuando los falsos negativos son muy costosos.
  Ventajas: adecuada en contextos donde la prioridad es no perder ninguna instancia positiva.
  Limitaciones: puede ser elevada a expensas de una baja precisión; no considera los falsos positivos.

- **Puntuación F1 (F1-score)**: es la media armónica de la precisión y la recuperación, definida como
  (F1 = 2 * (Precision * Recall) / (Precision + Recall)).
  Interpretación: combina precisión y recuperación en una sola métrica balanceada; alcanza su máximo cuando ambas métricas son altas.
  Ventajas: útil cuando se necesita un equilibrio entre precisión y recuperación; apta para conjuntos desequilibrados.
  Limitaciones: no considera los verdaderos negativos; puede enmascarar problemas si una clase es muy dominante.

- **AUC-ROC**: la curva ROC traza la tasa de verdaderos positivos frente a la tasa de falsos positivos a distintos umbrales. El área bajo la curva (AUC) representa la probabilidad de que el modelo asigne una puntuación más alta a un ejemplo positivo que a uno negativo.
  Interpretación: un AUC de 1 indica un modelo perfecto y 0.5 equivale a clasificar al azar.

Ventajas: permite comparar modelos independientemente del umbral de decisión; sintetiza la sensibilidad y la especificidad.

Limitaciones: en conjuntos muy desequilibrados puede no reflejar adecuadamente el rendimiento; la curva puede ser difícil de interpretar para usuarios no técnicos.

2.2 Métricas de regresión

- **Error absoluto medio (MAE)**: es la media de los valores absolutos de las diferencias entre predicciones y valores reales. Se calcula como
  ($MAE = (1/n) \times \Sigma|Actual - Predicted|$).
  Interpretación: mide el tamaño promedio de los errores sin considerar su signo. Mantiene la misma escala que los datos originales.
  Ventajas: fácil de interpretar; cada error contribuye de forma lineal, por lo que no penaliza excesivamente los grandes errores.
  Limitaciones: al no elevar al cuadrado las diferencias, no distingue entre errores grandes y pequeños; no es derivable matemáticamente.

- **Raíz del error cuadrático medio (RMSE)**: es la raíz cuadrada de la media de los errores al cuadrado. Su fórmula es
  ($RMSE = \sqrt{(\Sigma (Pi - Oi)^2 / n)}$).
  Interpretación: al elevar al cuadrado los errores antes de promediarlos, penaliza más los errores grandes. Valores más bajos indican mejor ajuste.
  Ventajas: sensible a las desviaciones grandes; mantiene las unidades originales del objetivo.
  Limitaciones: muy afectado por valores atípicos; no informa por sí solo del sesgo de las predicciones.

3. Solución con K-Nearest Neighbors (KNN)

3.1 Preparación de los datos

Se contó con una matriz de datos con dos variables predictoras: **glucosa** (niveles de glucosa en sangre) y **edad**, además de una **etiqueta** binaria que indica la presencia (1) o ausencia (0) de una condición. Para fines ilustrativos se generó un conjunto de 200 observaciones con una distribución equilibrada de la etiqueta. Se dividieron los datos en

un conjunto de entrenamiento (70 %) y un conjunto de prueba (30 %) utilizando estratificación para mantener la proporción de clases. Debido a que KNN se basa en distancias, ambas variables se escalaron mediante la estandarización (media cero y desviación unitaria) antes del entrenamiento.

3.2 Entrenamiento y selección de k

Se entrenaron clasificadores KNN con tres valores diferentes de *k*: 3, 5 y 7. Para cada modelo se calculó el F1-score en el conjunto de prueba. El valor de *k* que obtuvo el mejor F1-score fue **7**, lo que sugiere que considerar más vecinos condujo a un mejor equilibrio entre precisión y recuperación en este conjunto.

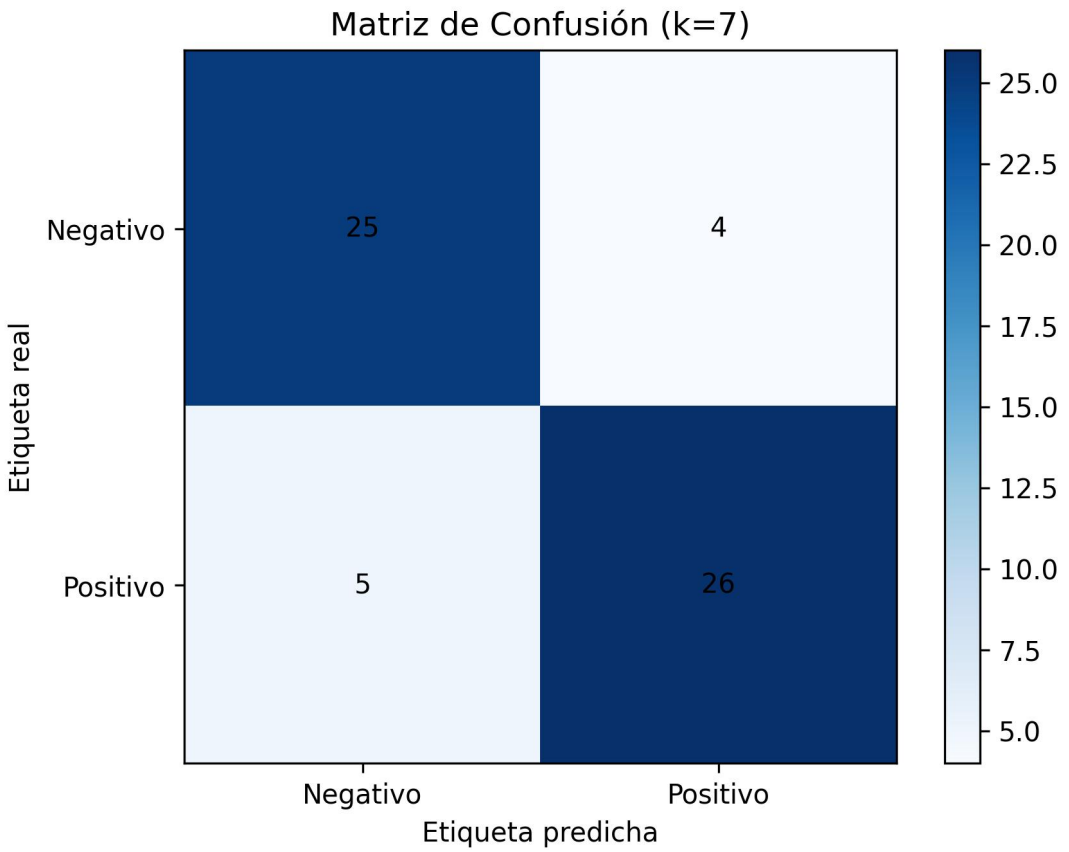
3.3 Evaluación del modelo

Se calcularon las métricas de exactitud, precisión, recuperación, F1 y AUC para cada valor de *k*. La Tabla 1 resume el F1-score y el AUC obtenidos:

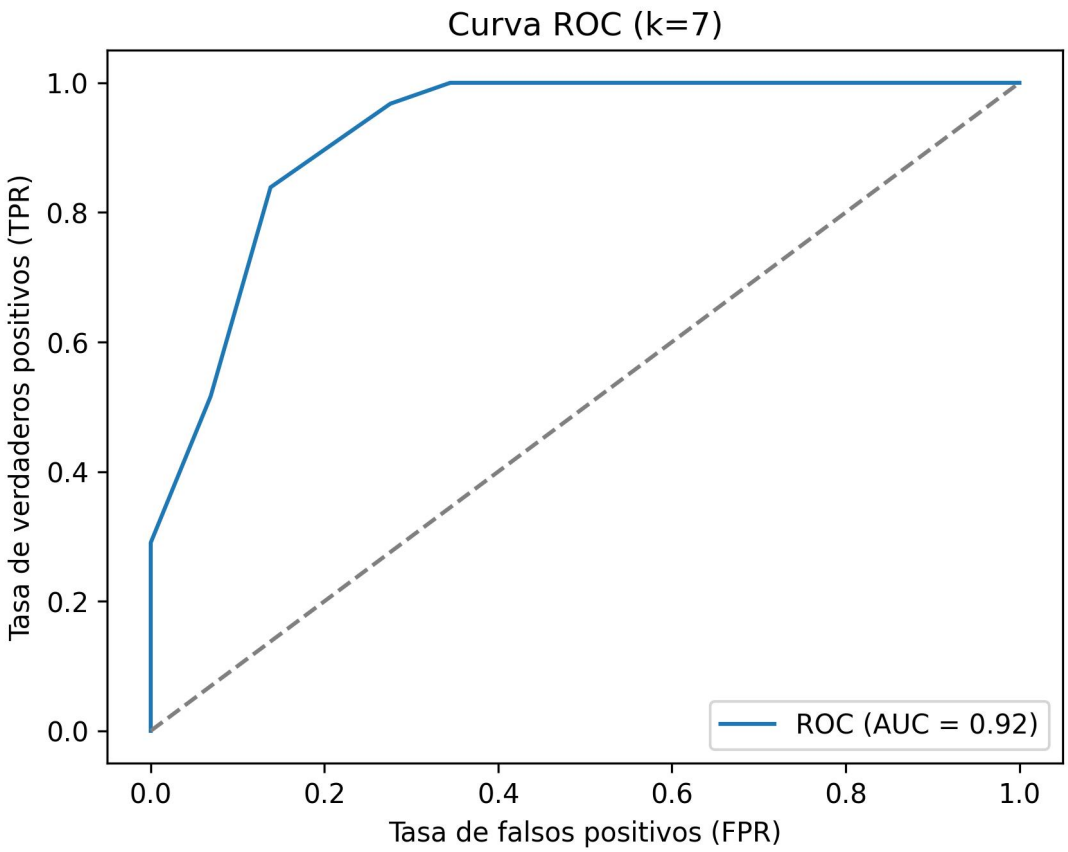
Valor de k	F1-score	AUC
3	0.81	0.90
5	0.82	0.90
7	**0.85**	**0.92**

El modelo con $k = 7$ logró una exactitud del 85 % y un F1-score de 0.85, con una precisión de 0.87 y una recuperación de 0.84. Su AUC de 0.92 indica un buen poder discriminativo.

A continuación se muestran la matriz de confusión y la curva ROC para el modelo seleccionado:



Matriz de confusión



Curva ROC

4. Resultados

La matriz de confusión muestra que el modelo con $k = 7$ clasificó correctamente 25 casos negativos y 26 casos positivos, con 4 falsos positivos y 5 falsos negativos. La curva ROC correspondiente presenta un área bajo la curva de **0.92**, lo que confirma el buen desempeño del clasificador al distinguir entre ambas clases a través de distintos umbrales de decisión.

Análisis de resultados

- **Comparación de k**: aunque los tres valores de k evaluados ofrecieron una exactitud similar, el aumento de k mejoró de manera consistente el F1-score y el AUC. Esto sugiere que, para estos datos, un número mayor de vecinos suaviza la influencia de ruido y produce predicciones más robustas.

- **Eficacia del modelo**: el modelo seleccionado alcanza un equilibrio adecuado entre precisión y recuperación. La presencia de algunos falsos negativos indica que todavía se podrían mejorar la detección de la clase positiva.
- **Posibles mejoras**:
  - Utilizar validación cruzada y una búsqueda en rejilla para explorar un rango más amplio de valores de k u otros hiperparámetros como el tipo de distancia.
  - Probar algoritmos alternativos (por ejemplo, regresión logística, árboles de decisión o máquinas de soporte vectorial) que puedan capturar relaciones no lineales entre las variables.
  - Incorporar más variables predictoras relevantes o recolectar más datos para mejorar la generalización del modelo.
  - Aplicar técnicas de balanceo de clases si en un escenario real la etiqueta estuviera desbalanceada.

5. Conclusiones y recomendaciones

El estudio de métricas de clasificación y regresión permite seleccionar criterios adecuados para evaluar modelos. La exactitud es útil como referencia general, pero métricas como precisión, recuperación y F1-score ofrecen información más detallada cuando existen desequilibrios de clase o distintas penalizaciones para los errores. La AUC-ROC proporciona un resumen independiente del umbral y facilita la comparación entre modelos.

En la aplicación práctica, el clasificador KNN demostró ser una opción efectiva para predecir la etiqueta binaria utilizando las variables de glucosa y edad. Con un valor de k igual a 7 se obtuvo el mejor equilibrio entre sensibilidad y especificidad, con un AUC elevado. Para mejorar el rendimiento se recomienda experimentar con otros algoritmos y validar sistemáticamente los hiperparámetros.

6. Referencias

- Google Developers. (2025). *Clasificación: Exactitud, recuperación, precisión y métricas relacionadas*. Recuperado de https://developers.google.com/machine-learning/crash-course/classification/accuracy-precision-recall

- Google Developers. (2025). *Clasificación: ROC y AUC*. Recuperado de https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc

- Wikipedia. (2025). *Mean absolute error*. Recuperado de https://en.wikipedia.org/wiki/Mean_absolute_error

- Deepchecks. (s.f.). *Root Mean Square Error (RMSE)*. Recuperado de https://www.deepchecks.com/glossary/root-mean-square-error/