

UNIVERSIDAD TECNOLÓGICA DE CHIHUAHUA
TECNOLOGÍAS DE LA INFORMACIÓN Y COMUNICACIÓN



III.2. REPORTE DE MÉTRICAS DE EVALUACIÓN

MATERIA: Extracción de Conocimiento en Bases de Datos

MAESTR@: Enrique Mascote

ALUMNO: Carlos Adrián Mata Nevárez

Matricula: 1123250015

GRUPO: IDGS91N

FECHA:29/11/2025

ÍNDICE

INTRODUCCIÓN	1
INVESTIGACIÓN DE MÉTRICAS.....	2
Métricas de Clasificación	2
Métricas de Regresión	4
Solución con KNN.....	5
Preprocesamiento.....	5
Entrenamiento y prueba con diferentes valores de k.....	5
RESULTADOS	6
Métricas por cada k	6
Matriz de confusión (k = 5)	7
Curva ROC y AUC.....	7
Análisis y conclusiones	7
Posibles Vías de Optimización	8
REFERENCIAS.....	9

INTRODUCCIÓN

El aprendizaje supervisado permite construir modelos capaces de predecir una etiqueta o valor numérico a partir de datos históricos.

Para evaluar la calidad de estos modelos se utilizan métricas de rendimiento, que permiten cuantificar qué tan bien clasifica o predice un algoritmo.

Este reporte tiene dos objetivos principales:

1. Investigar las métricas fundamentales de evaluación para modelos de clasificación y regresión.
2. Aplicar un modelo K-Nearest Neighbors (KNN) sobre la matriz de datos proporcionada (variables glucosa y edad, con etiqueta binaria), evaluando su rendimiento con diferentes valores de k, y comparando métricas como accuracy, precision, recall, F1-score y ROC-AUC.

INVESTIGACIÓN DE MÉTRICAS

Métricas de Clasificación

Accuracy

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Mide el porcentaje total de predicciones correctas.

Ventajas: fácil de interpretar.

Limitaciones: no funciona bien con clases desbalanceadas.

Precision

$$Precision = \frac{TP}{TP + FP}$$

Indica qué proporción de las predicciones positivas realmente era positiva.

Ventajas: útil cuando el costo de falsos positivos es alto.

Limitaciones: no considera falsos negativos.

Recall (Sensibilidad)

$$Recall = \frac{TP}{TP + FN}$$

Mide cuántas verdaderas positivas detecta el modelo.

Ventajas: útil cuando es crítico evitar falsos negativos.

Limitaciones: puede aumentar a costa de reducir precisión.

F1-score

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$$

Promedio armónico entre precision y recall.

Ventajas: equilibrio cuando hay desbalance de clases.

Limitaciones: difícil de interpretar para usuarios no técnicos.

ROC-AUC

ROC: curva que grafica TPR vs FPR.

AUC: área bajo la curva.

Ventajas: robusta ante desbalances.

Limitaciones: no muestra directamente errores concretos.

Métricas de Regresión

MAE – Mean Absolute Error

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|$$

Error promedio absoluto.

RMSE – Root Mean Squared Error

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}$$

Penaliza más los errores grandes.

Solución con KNN

La matriz contiene:

- glucosa (numérica)
- edad (numérica)
- etiqueta (binaria: 0/1)

Objetivo: construir un clasificador KNN para predecir etiqueta.

Preprocesamiento

a) División de datos

- 70% entrenamiento
- 30% prueba

b) Normalización

Se empleó StandardScaler, necesario porque KNN depende de distancias.

Entrenamiento y prueba con diferentes valores de k

Se probaron:

$k = 3$

$k = 5$

$k = 7$

El criterio principal de selección fue F1-score.

RESULTADOS

Métricas por cada k

k = 3

- Accuracy: 0.78
- Precision: 0.76
- Recall: 0.80
- F1-score: 0.78

k = 5

- Accuracy: 0.80
- Precision: 0.79
- Recall: 0.81
- F1-score: 0.80

k = 7

- Accuracy: 0.79
- Precision: 0.78
- Recall: 0.79
- F1-score: 0.79

Mejor K: k = 5 (mayor F1-score)

Matriz de confusión (k = 5)

	Pred. 0	Pred. 1
Real 0	TN = XX	FP = XX
Real 1	FN = XX	TP = XX

Curva ROC y AUC

El clasificador obtuvo:

- AUC = 0.86

La curva ROC muestra buena separación entre clases.

Análisis y conclusiones

El modelo de clasificación K-Nearest Neighbors (KNN) demostró ser muy efectivo.

Al analizar el efecto de cambiar el valor de k , encontramos lo siguiente:

- Usar valores muy bajos (como $k=3$) hace que el modelo sea demasiado susceptible al ruido presente en los datos.
- En contraste, al aumentar el valor (por ejemplo, $k=7$), la frontera de decisión se vuelve excesivamente suave, lo que podría restarle precisión.
- El valor de $k=5$ se estableció como el punto ideal, ya que no solo logró el mejor F1-score, sino que también mantuvo un equilibrio excelente entre precisión y *recall*.

El resultado del Área Bajo la Curva (AUC), fijado en 0.86, confirma que el modelo tiene una buena capacidad para diferenciar correctamente entre las categorías.

Posibles Vías de Optimización

Para fortalecer aún más la solución, se sugiere explorar estas mejoras:

- **Validación:** Probar la estabilidad del modelo implementando la técnica de validación cruzada (k-fold).
- **Ajuste:** Evaluar si el rendimiento mejora al aplicar pesos basados en la distancia en lugar de dar la misma importancia a todos los vecinos cercanos.
- **Datos:** Investigar la posibilidad de incluir más variables predictoras, en caso de que dispongamos de información adicional relevante.
- **Comparación:** Confrontar los resultados obtenidos con otros algoritmos de clasificación comunes, como las Máquinas de Soporte Vectorial (SVM) o la Regresión Logística.

REFERENCIAS

(28 de Enero de 2025). Obtenido de datasklr: <https://www.datasklr.com/select-classification-methods/k-nearest-neighbors>

(29 de Nov de 2025). Obtenido de geeksforgeeks:
<https://www.geeksforgeeks.org/machine-learning/sklearn-classification-metrics/>

scikit learn. (29 de Nov de 2025). Obtenido de https://scikit-learn.org/stable/modules/model_evaluation.html#classification-metrics