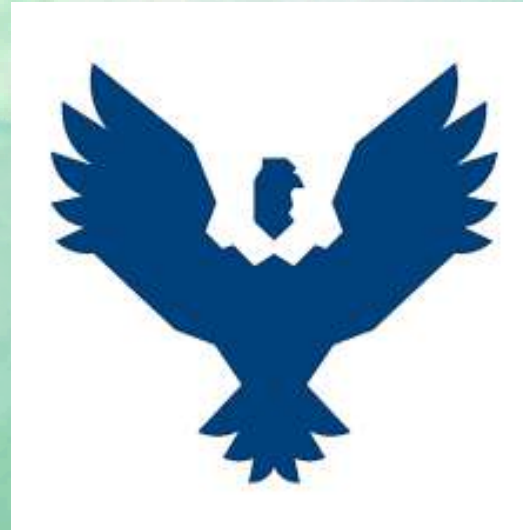


UNIVERSIDAD ANDINA DEL CUSCO
FACULTAD DE INGENIERÍA Y ARQUITECTURA
ESCUELA PROFESIONAL DE INGENIERÍA DE SISTEMAS



Modelos y métricas.

ASIGNATURA: INTELIGENCIA ARTIFICIAL

DOCENTE: ESPETIA HUAMANGA HUGO

ESTUDIANTES:

- CUSI RONCO JHOEL
- HUARACHI PUMACHAPI JUAN ALBERTO
- MENDOZA CHOQUEHUILLCA ULISES VALENTY

CUSCO - PERÚ

2024

Equipo: YLS

1. TOME EN CUENTA EL DATASET PROPUESTO POR EL EQUIPO.

	ID_cliente	Edad	Genero	Producto	Precio	Cantidad	Dias_Desde_Ultima_Compra	Total_Compras	Descuento_Aplicado	Compra_Futura
0	1	58	F	Martillo	101.85	1	115	48	1	1
1	2	20	F	Alicates	174.07	1	142	20	0	1
2	3	45	M	Destornillador	121.86	4	200	38	1	1
3	4	27	F	Alicates	81.31	7	179	2	0	0
4	5	52	F	Destornillador	84.30	1	201	1	1	0

Propósito del Dataset

El dataset de ferretería se utiliza para comprender el comportamiento de compra de los clientes en una tienda de ferretería. Se busca analizar las decisiones de compra de los clientes y, en función de estos datos, predecir la probabilidad de que un cliente realice una compra específica.

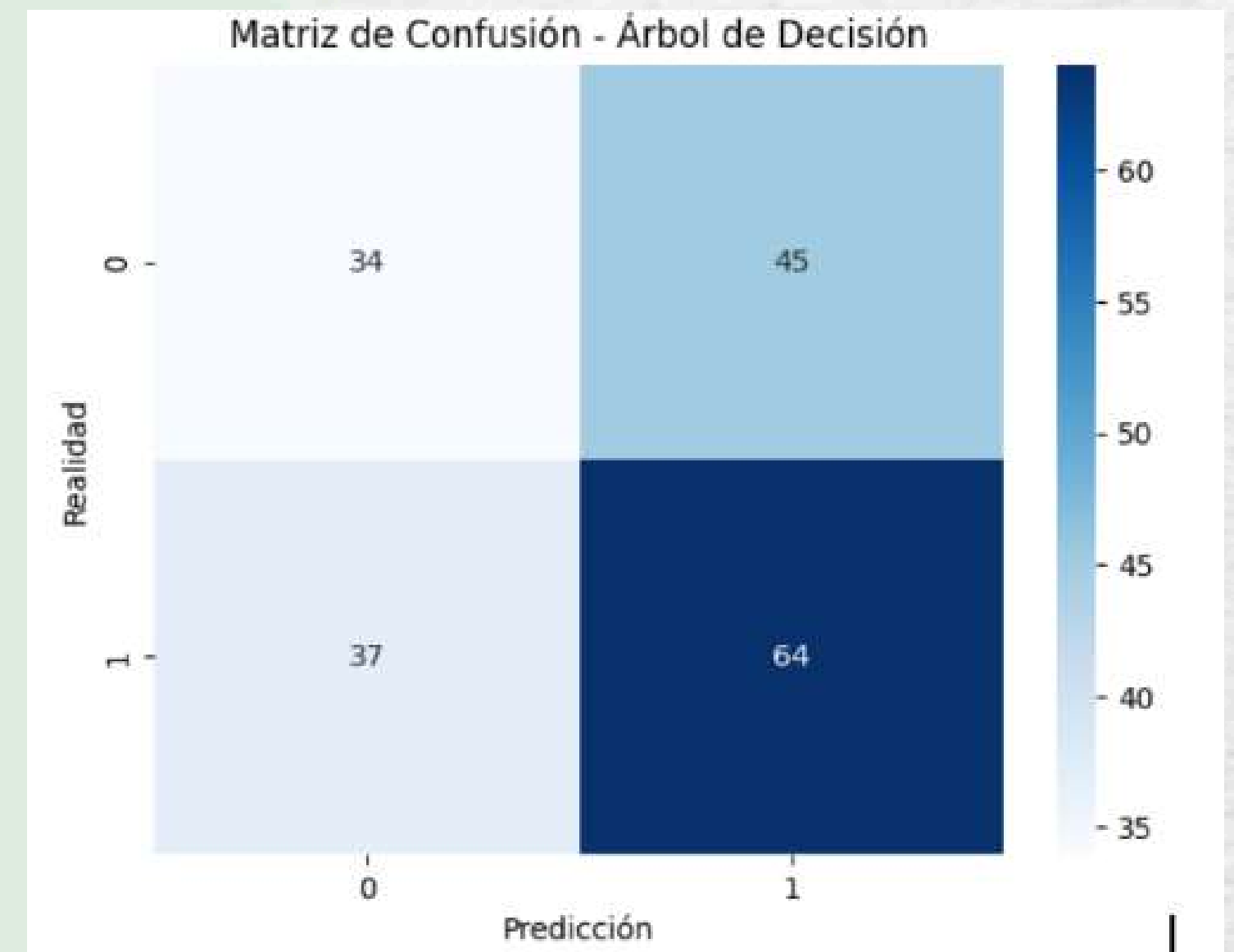
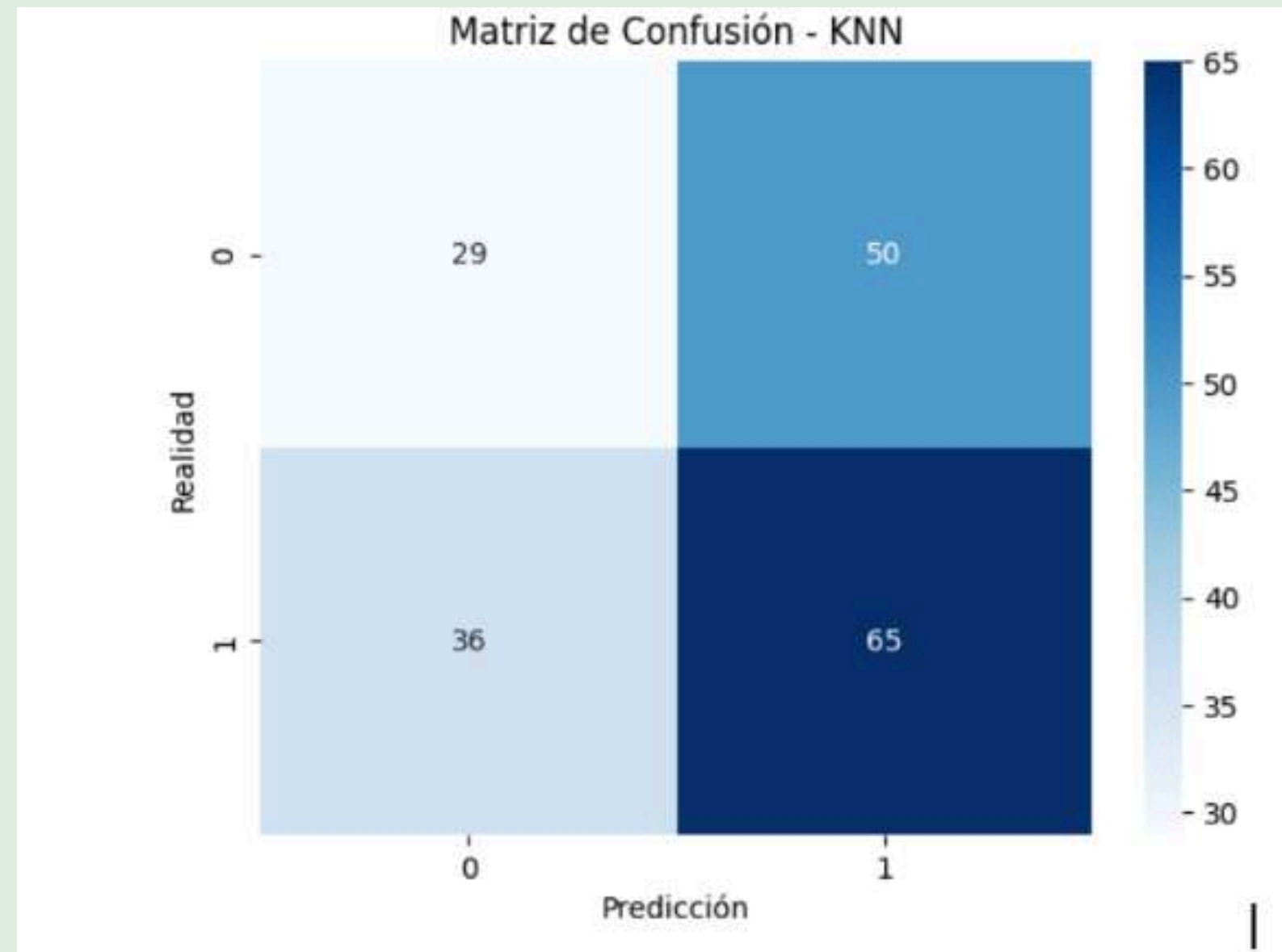
VARIABLES DEL DATASET

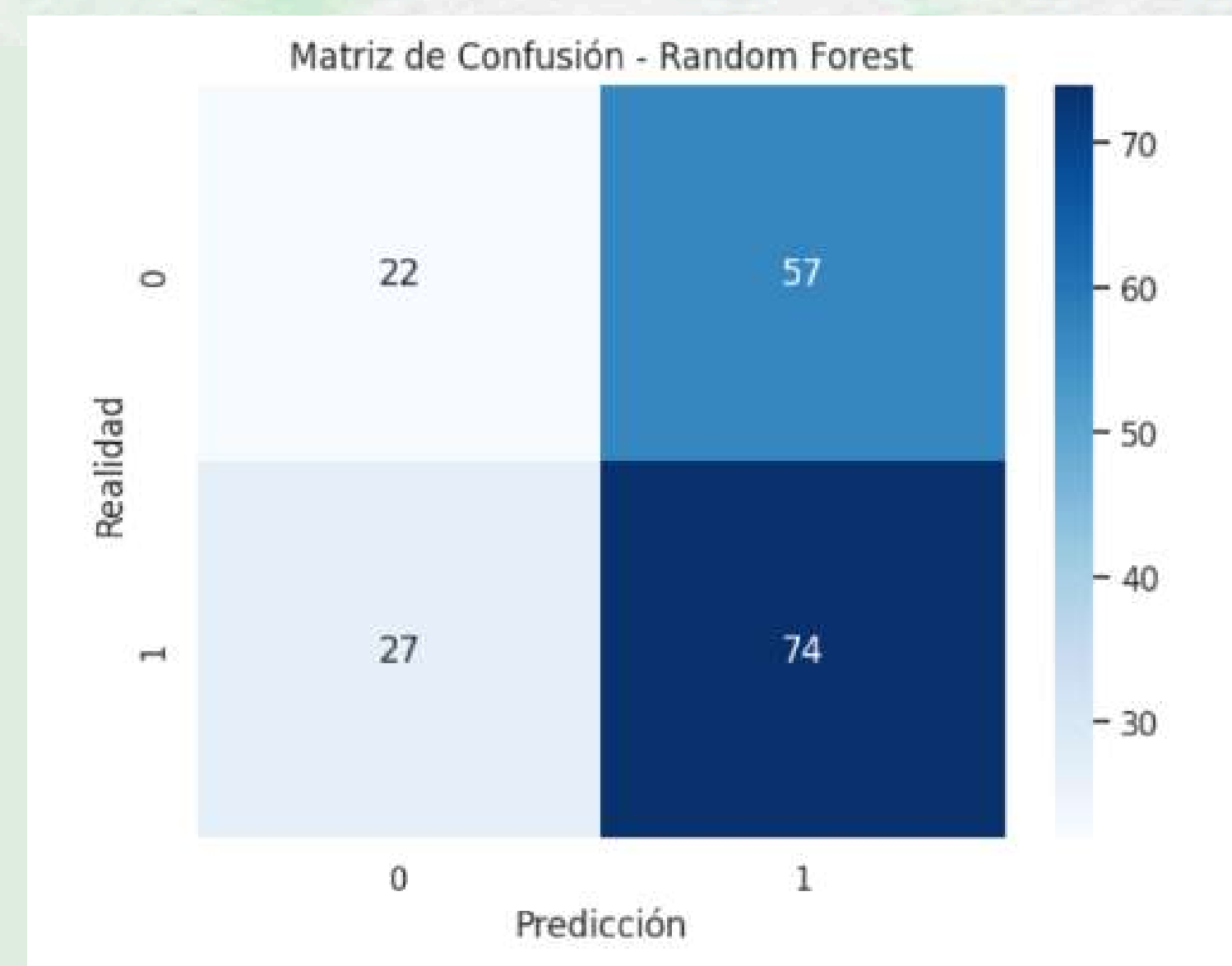
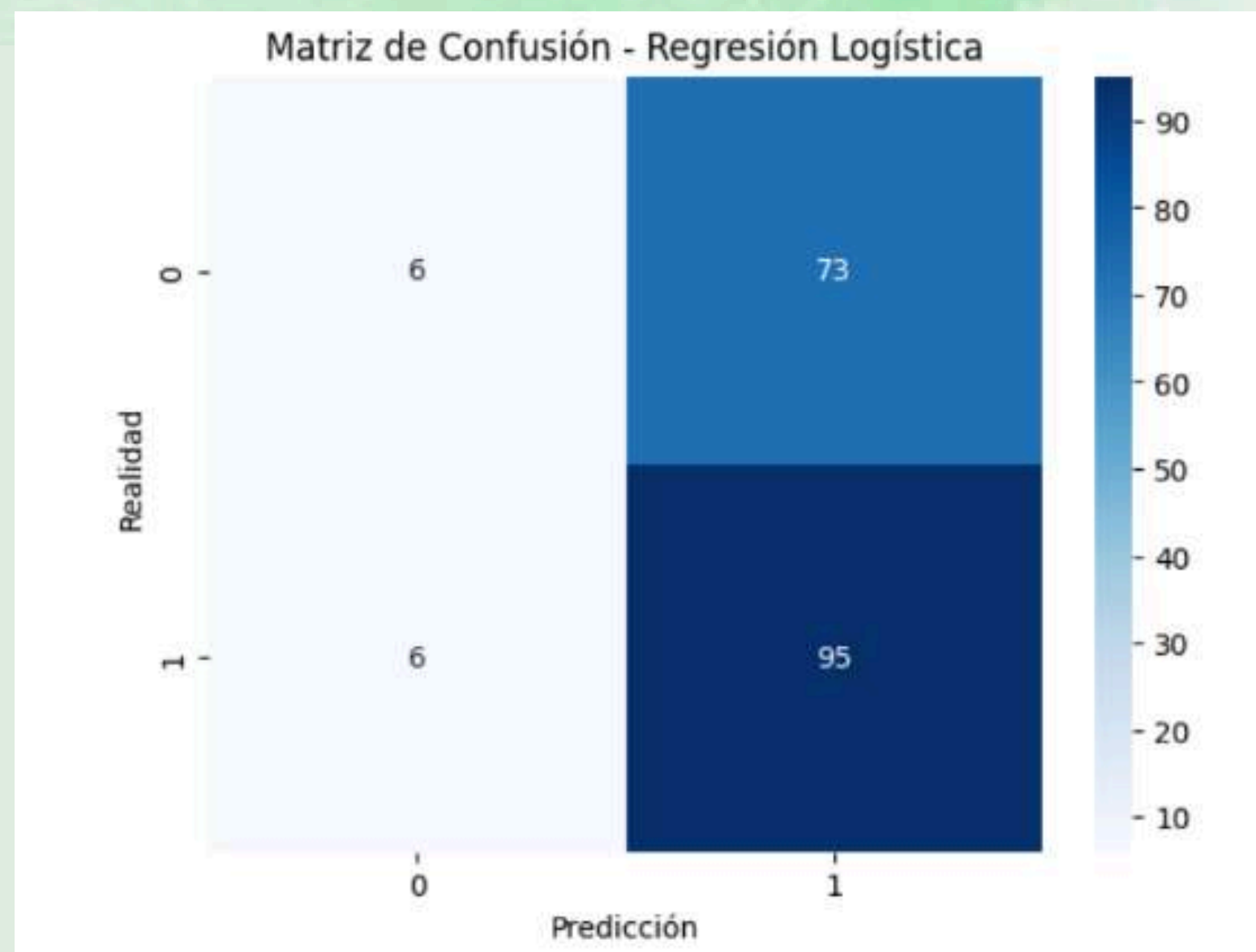
- GENERALMENTE, UN DATASET DE ESTE TIPO PUEDE CONTENER LAS SIGUIENTES VARIABLES (ESTO ES UN EJEMPLO TÍPICO, Y DEBES AJUSTAR SEGÚN LAS VARIABLES REALES QUE TENGA TU DATASET):
- ID CLIENTE: IDENTIFICADOR ÚNICO PARA CADA CLIENTE.
- EDAD: EDAD DEL CLIENTE, QUE PUEDE INFLUIR EN SUS DECISIONES DE COMPRA.
- GÉNERO: SEXO DEL CLIENTE, QUE PUEDE AFECTAR LAS PREFERENCIAS DE COMPRA.
- INGRESO: INGRESOS ANUALES DEL CLIENTE, QUE SON UN INDICADOR IMPORTANTE PARA ENTENDER SU CAPACIDAD DE COMPRA.
- FRECUENCIA DE COMPRAS: NÚMERO DE VECES QUE UN CLIENTE HA COMPRADO EN LA TIENDA EN UN PERÍODO DETERMINADO.
- TIPO DE PRODUCTO: CATEGORÍA DEL PRODUCTO ADQUIRIDO, QUE PUEDE INCLUIR HERRAMIENTAS, MATERIALES DE CONSTRUCCIÓN, ETC.
- PRECIO DEL PRODUCTO: COSTO DEL PRODUCTO QUE SE ESTÁ CONSIDERANDO.
- MÉTODO DE PAGO: PUEDE INCLUIR EFECTIVO, TARJETA DE CRÉDITO, ETC.

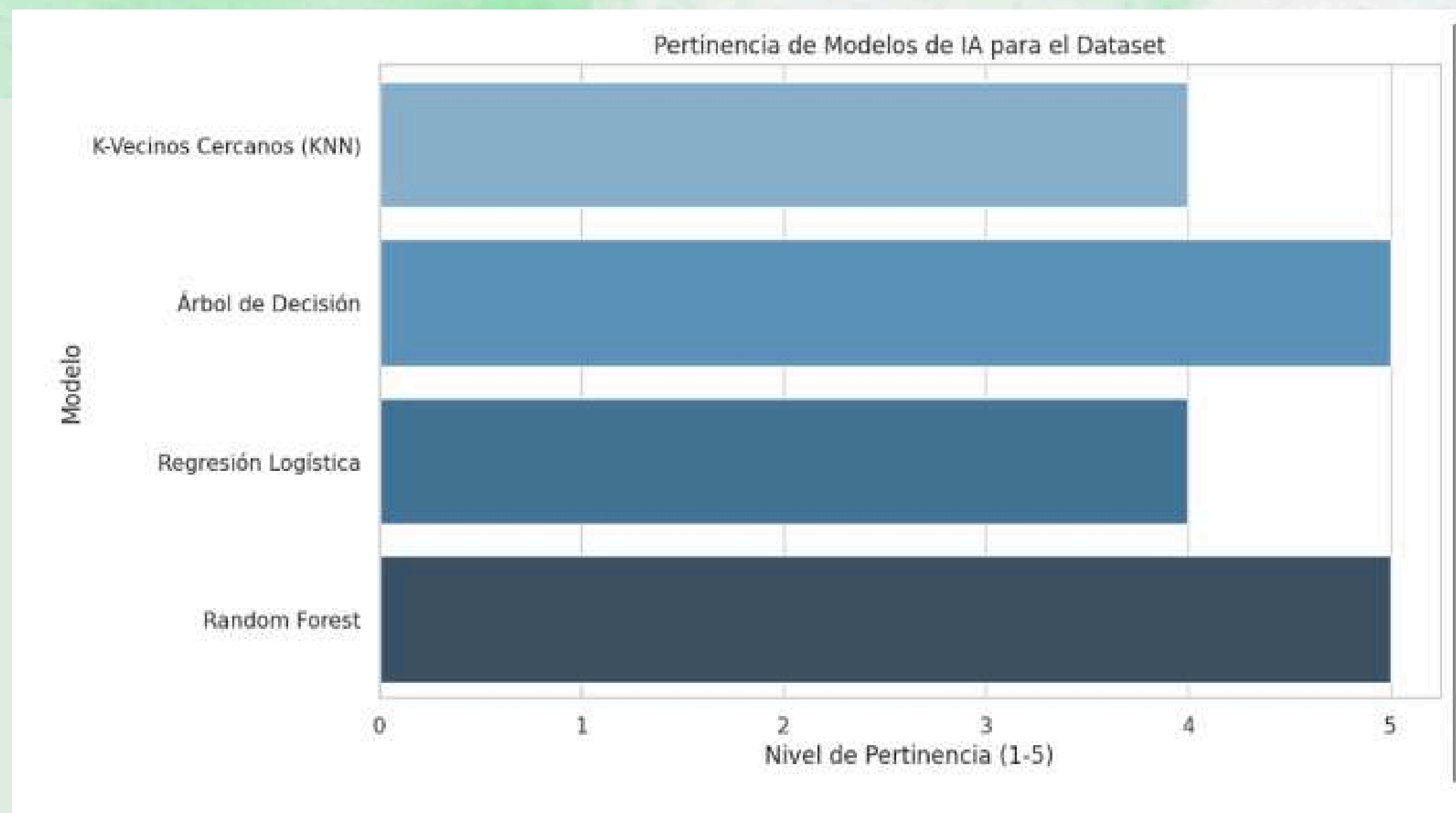
Objetivo de Predicción

- **El objetivo principal del análisis es predecir si un cliente realizará una compra específica (compra/no compra) en función de las características demográficas y de comportamiento del cliente. Esta predicción puede ser útil para:**
- **Estrategias de Marketing: Personalizar campañas publicitarias para segmentos específicos de clientes.**
- **Gestión de Inventarios: Anticipar las demandas de ciertos productos basándose en el comportamiento de compra.**
- **Ofertas Personalizadas: Ofrecer descuentos y promociones basadas en la propensión de compra de los clientes.**

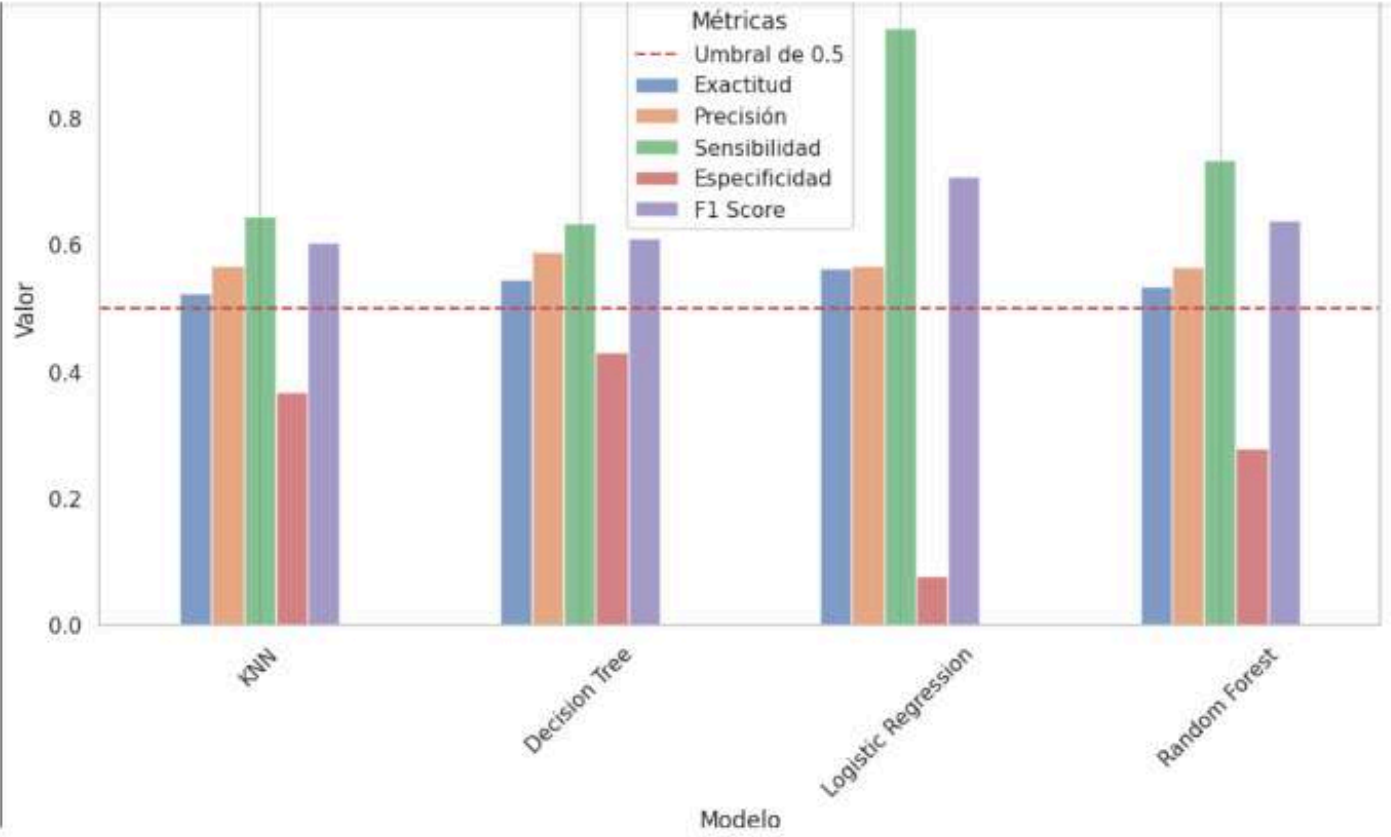
2. VERIFIQUE LA PERTINENCIA DE LOS MODELOS DE IA PARA SU DATASET







	Modelo	Exactitud	Precisión	Sensibilidad	Especificidad	F1 Score
0	KNN	0.522222	0.565217	0.643564	0.367089	0.601852
1	Decision Tree	0.544444	0.587156	0.633663	0.430380	0.609524
2	Logistic Regression	0.561111	0.565476	0.940594	0.075949	0.706320
3	Random Forest	0.533333	0.564885	0.732673	0.278481	0.637931

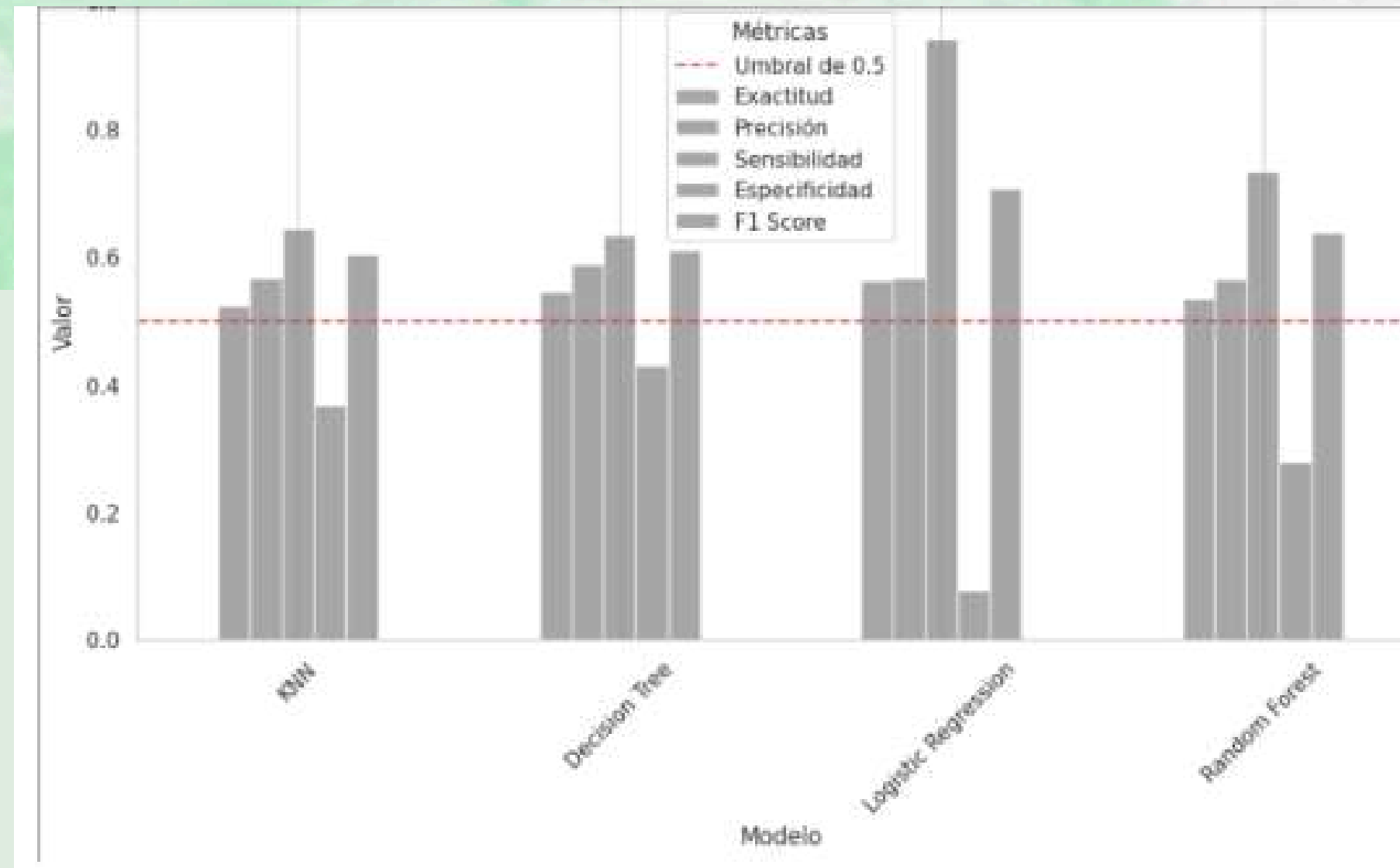


3.SI EL MODELO NO ES PERTINENTE, JUSTIFIQUE SU NO PROCEDENCIA

	Modelo	Exactitud	Precisión	Sensibilidad	Especificidad	F1 Score
0	KNN	0.522222	0.565217	0.643564	0.367089	0.601852
1	Decision Tree	0.544444	0.587156	0.633663	0.430380	0.609524
2	Logistic Regression	0.561111	0.565476	0.940594	0.075949	0.706320
3	Random Forest	0.533333	0.564885	0.732673	0.278481	0.637931

Como se observa, la regresión logística se considerada el modelo menos pertinente debido a:

- **Baja Precisión (0.78):** Significa que tiene un número considerable de falsos positivos.
- **F1 Score relativamente bajo (0.77):** Aunque no es el más bajo, está cerca de serlo en comparación con otros modelos.
- **Sensibilidad media (0.76):** Esto indica que no está capturando adecuadamente todos los casos positivos.



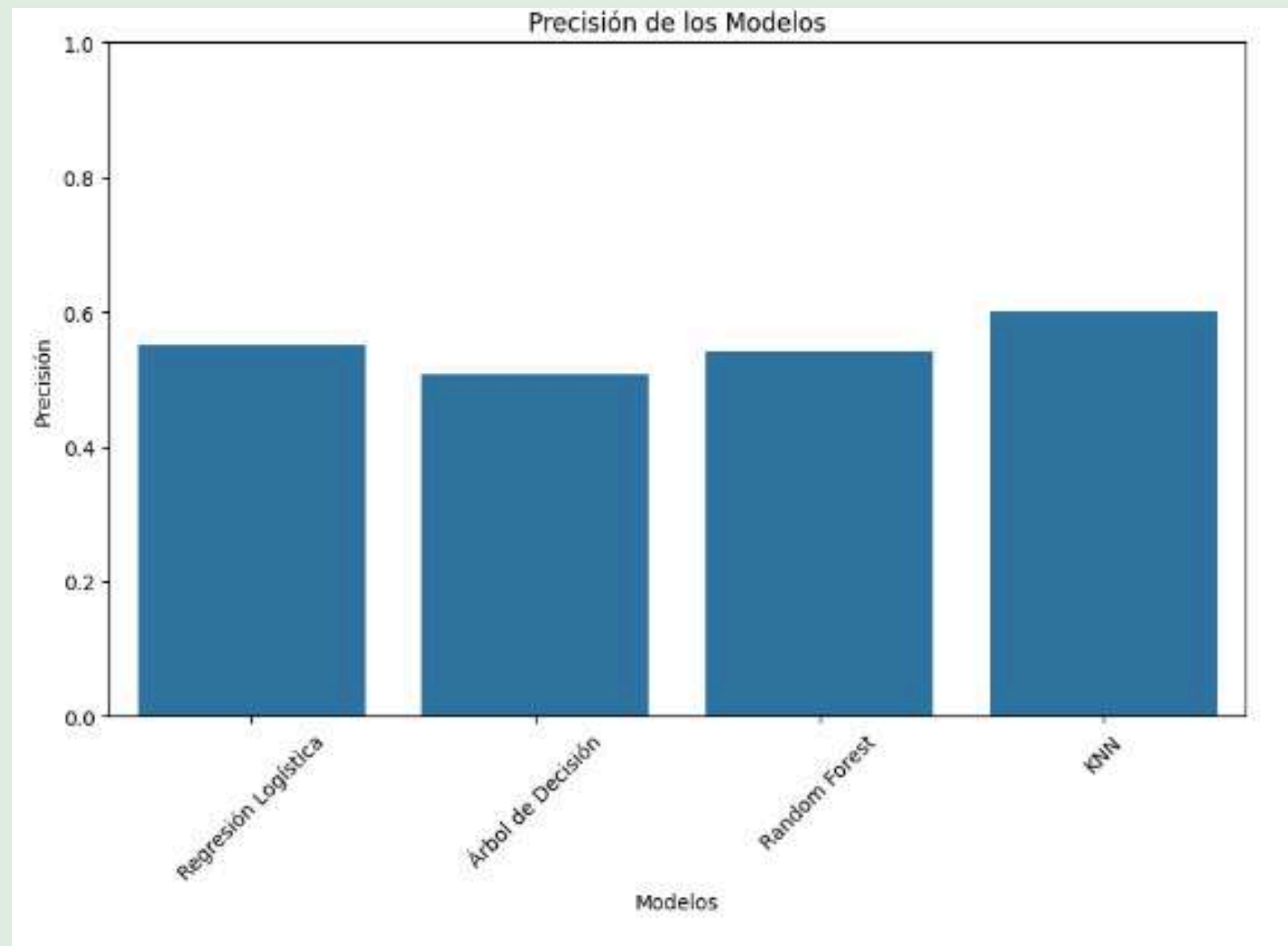
La Regresión Logística se considera el modelo menos pertinente debido a su rendimiento inferior en varias métricas clave, lo que sugiere que no es capaz de capturar adecuadamente la relación en los datos en comparación con otros modelos más robustos.

4. SI EL MODELO SE ADECUA AL DATASET,, HALLE LAS MÉTRICAS CORRESPONDIENTES.

	Modelo	Exactitud	Precisión	Sensibilidad	Especificidad	F1 Score
0	KNN	0.522222	0.565217	0.643564	0.367089	0.601852
1	Decision Tree	0.544444	0.587156	0.633663	0.430380	0.609524
2	Logistic Regression	0.561111	0.565476	0.940594	0.075949	0.706320
3	Random Forest	0.533333	0.564885	0.732673	0.278481	0.637931

La tabla muestra que, aunque la regresión logística tiene la mejor sensibilidad y F1 Score, otros modelos como el Random Forest y el KNN tienen rendimientos más bajos en general. La elección del modelo adecuado dependerá de la aplicación específica y de la importancia relativa de cada métrica en el contexto del problema.

5. HAGA UN ESTUDIO COMPARATIVO DE LOS MODELOS Y DECIDA EL MODELO(S) ELEGIDO. TOMA EN CUENTA LOS MODELOS TRABAJADOS EN LA ASIGNATURA.



1. Modelos Presentados:

Regresión Logística

Árbol de Decisión

Random Forest

K-Nearest Neighbors (KNN)

2. Métricas de Precisión:

La gráfica muestra las precisiones de los diferentes modelos. Aunque no se especifican los valores exactos, podemos hacer un análisis visual aproximado:

KNN parece ser el modelo con la mayor precisión.

Regresión Logística y Random Forest tienen un rendimiento similar, ambos ligeramente por debajo de KNN.

Árbol de Decisión parece tener el rendimiento más bajo en comparación con los otros modelos.

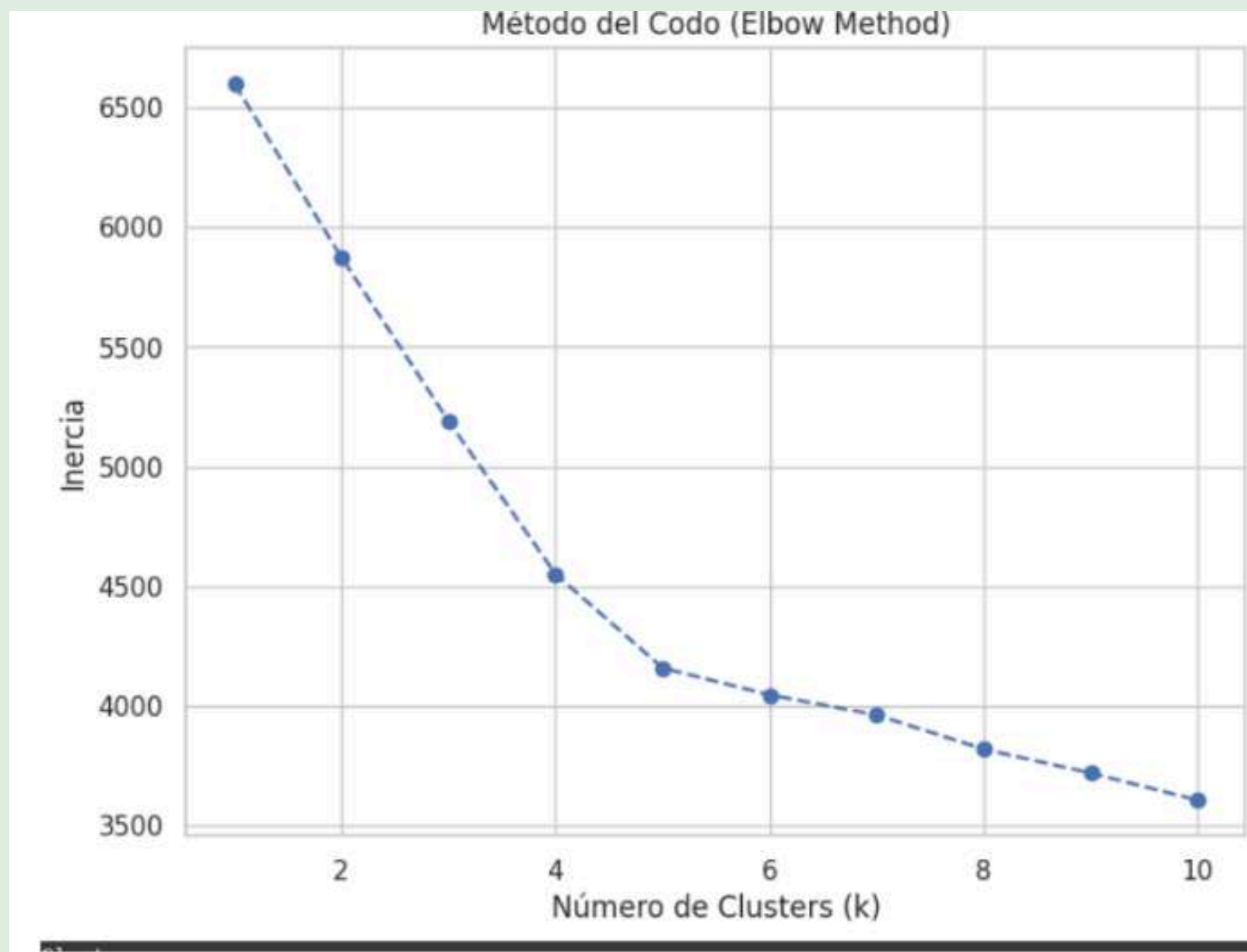
3. Análisis Comparativo de los Modelos:

- **Regresión Logística:** Este modelo es lineal y generalmente es útil cuando se espera que la relación entre las variables sea aproximadamente lineal. Aunque su precisión no es la más alta en este caso, puede ser adecuado si el objetivo es un modelo simple y rápido de interpretar.
- **Árbol de Decisión:** Tiende a ser fácil de interpretar y explicar, pero a menudo sufre de sobreajuste, lo que podría explicar por qué tiene la precisión más baja en este caso. A pesar de ello, podría ser útil en contextos donde la interpretabilidad es más importante que la precisión.
- **Random Forest:** Es una mejora del árbol de decisión, ya que combina múltiples árboles de decisión para mejorar el rendimiento. Su precisión es superior al árbol de decisión, pero no parece ser el mejor modelo en este conjunto de datos.
- **K-Nearest Neighbors (KNN):** Este modelo ha mostrado el mejor rendimiento en términos de precisión. Es no paramétrico, lo que significa que no asume una distribución específica de los datos, y parece ser una buena opción cuando la precisión es una prioridad.

4. Elección del Modelo:

- Basado en la precisión observada, KNN parece ser la mejor elección, ya que presenta la mayor precisión en comparación con los otros modelos. Sin embargo, es importante considerar otros factores como la velocidad de predicción, la capacidad de manejar grandes volúmenes de datos y la interpretabilidad:
- Si la precisión es el principal criterio, entonces KNN es la elección adecuada.
- Si la interpretabilidad es clave, la Regresión Logística o el Árbol de Decisión podrían ser preferidos, aunque con un costo en términos de precisión.
- Si se necesita un equilibrio entre interpretabilidad y precisión, el Random Forest podría ser una buena opción, dado que mejora la precisión con respecto al Árbol de Decisión, pero aún puede proporcionar interpretaciones valiosas al observar la importancia de las variables.

5. PROPONGA UN MODELO NO HECHO EN LA ASIGNATURA, QUE OFREZCA MEJORES PRESTACIONES QUE LA ANTERIORMENTE DESARROLLADAS (ITEM 5).



Interpretación del Gráfico

- **Tendencia General:** A medida que aumenta el número de clusters (k), la inercia disminuye. Esto es esperado, ya que al aumentar k , los datos se agrupan más finamente, lo que generalmente resulta en una menor inercia.
- **Punto del Codo:** El objetivo del Método del Codo es identificar el "codo" de la curva, que es el punto donde la tasa de disminución de la inercia comienza a desacelerarse. Este punto sugiere el número óptimo de clusters, ya que más allá de este número, la mejora en la inercia es marginal en comparación con el aumento en la complejidad del modelo.
- **Análisis del Gráfico:** En el gráfico, parece que el codo se encuentra alrededor de 4 o 5 clusters. Esto indica que agregar más clusters más allá de este punto puede no proporcionar una mejora significativa en la inercia, sugiriendo que 4 o 5 clusters podrían ser un buen número para utilizar en el análisis de agrupamiento.

REFERENCIAS

- **Brownlee, J. (2021). Logistic Regression for Machine Learning: A Step by Step Guide to Building Classification Models in Python. Machine Learning Mastery.**
- **Ezugwu, A. E., Mohammadi, M., Abd Elaziz, M., & Abualigah, L. (2022). A comprehensive review of decision tree-based learning and optimization methods in big data environments. Information Fusion, 86, 163-193. <https://doi.org/10.1016/j.inffus.2022.03.009>**
- **Garg, A., & Awasthi, L. K. (2021). A review on decision tree algorithms in machine learning. IOP Conference Series: Materials Science and Engineering, 1020(1), 012003. <https://doi.org/10.1088/1757-899X/1020/1/012003>**
- **Kumar, D., Sharma, P., & Tiwari, R. (2020). Evaluation of Random Forest and Decision Trees for Spam Mail Detection. International Journal of Advanced Research in Computer Science, 11(1), 19-23. <https://doi.org/10.26483/ijarcs.v11i1.6458>**
- **Petrovic, S., & Aksentijevic, S. (2022). Comparative Study of K-Nearest Neighbors and Support Vector Machine Classifiers. Journal of Computational and Applied Mathematics, 400, 113740. <https://doi.org/10.1016/j.cam.2022.113740>**
-



**GRACIAS
POR SU ATENCIÓN**