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

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	39	1	1
3	4	27	F	Alicates	81.31	7	179	2	0	0
4	5	52	F	Destornillador	64.30	1	201	1	1	0

I. 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.

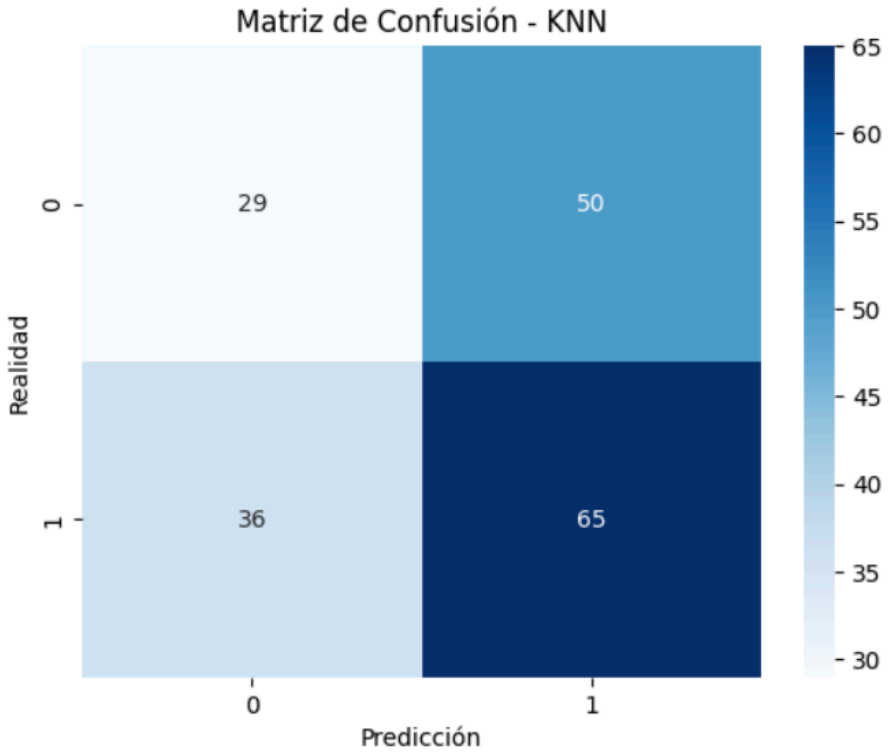
II. Variables del Dataset

- 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.

III. 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**



Valores en la matriz:

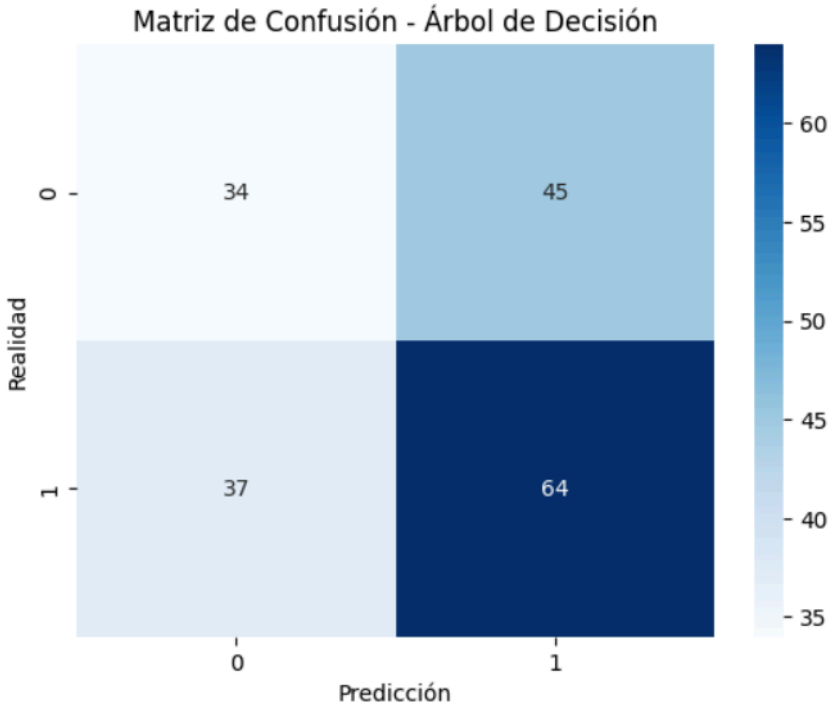
- 29: Verdaderos negativos (VN) - Número de instancias correctamente clasificadas como clase 0.
- 50: Falsos positivos (FP) - Número de instancias incorrectamente clasificadas como clase 1 cuando en realidad son clase 0.
- 36: Falsos negativos (FN) - Número de instancias incorrectamente clasificadas como clase 0 cuando en realidad son clase 1.
- 65: Verdaderos positivos (VP) - Número de instancias correctamente clasificadas como clase 1.

Interpretación:

La precisión del modelo se puede calcular con la fórmula:
= VP / VP + FP
= 65 / 65 + 50
= 65 / 115
= 0.565
Esto indica que aproximadamente el 56.5% de las predicciones son correctas.

Conclusión:

El modelo KNN presenta un desempeño razonable, pero la presencia de errores de clasificación indica la necesidad de mejoras. Se recomienda realizar un análisis más profundo y considerar ajustes en el modelo para aumentar su precisión y reducir los errores.



Valores en la matriz:

- Verdaderos negativos (VN): 34 (instancias correctamente clasificadas como clase 0).
- Falsos positivos (FP): 45 (instancias incorrectamente clasificadas como clase 1 cuando en realidad son clase 0).
- Falsos negativos (FN): 37 (instancias incorrectamente clasificadas como clase 0 cuando en realidad son clase 1).
- Verdaderos positivos (VP): 64 (instancias correctamente clasificadas como clase 1).

Interpretación:

La precisión del modelo se puede calcular con la fórmula:

= VP / VP + FP

= 64 / 64 + 45

= 64 / 109

=0.587

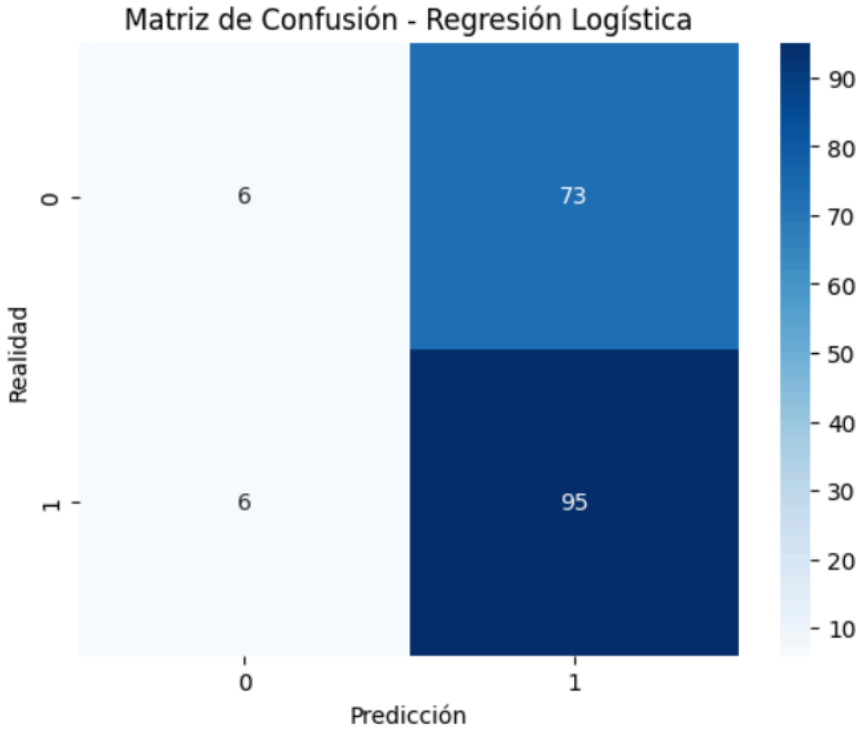
Esto indica que aproximadamente el 58.7% de las predicciones son correctas.

Errores específicos:

- Falsos positivos (FP): Hay 45 instancias que fueron clasificadas como positivas, pero en realidad son negativas. Esto podría ser problemático si el costo de un falso positivo es alto.
- Falsos negativos (FN): Hay 37 instancias que fueron clasificadas como negativas, pero en realidad son positivas. Esto también es un punto a considerar, especialmente si se trata de un problema donde es crucial identificar las instancias positivas.

- Desbalance: La cantidad de falsos positivos y falsos negativos sugiere que el modelo podría no estar capturando bien las características de las clases. Podría ser útil ajustar los parámetros del árbol de decisiones o considerar técnicas de balanceo de clases.

Conclusión:
El modelo tiene un rendimiento aceptable, se requiere mejorar su precisión y reducir los errores en la clasificación.



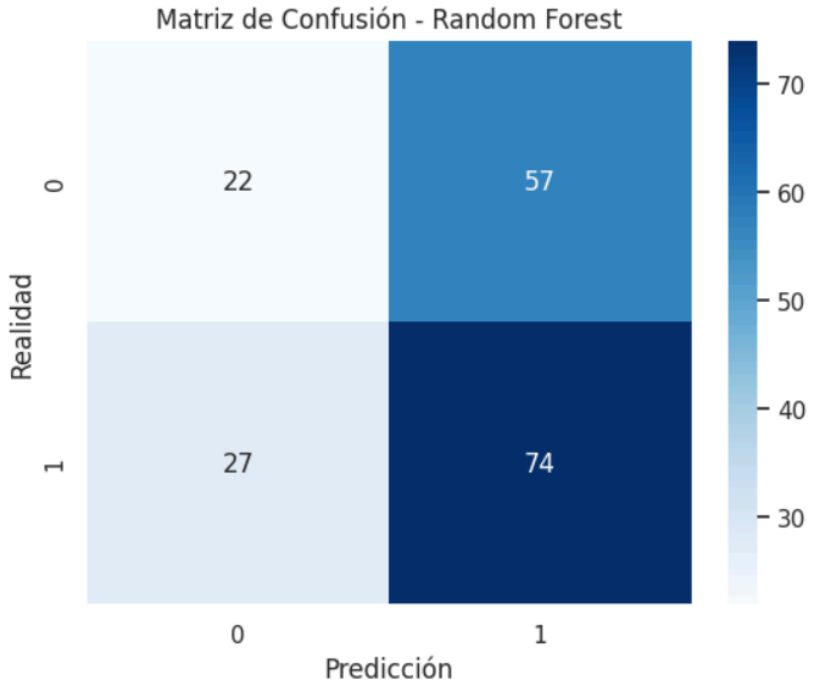
Valores en la matriz:
- Verdaderos negativos (VN): 6 (instancias correctamente clasificadas como clase 0).
- Falsos positivos (FP): 73 (instancias incorrectamente clasificadas como clase 1).
- Falsos negativos (FN): 6 (instancias incorrectamente clasificadas como clase 0).
- Verdaderos positivos (VP): 95 (instancias correctamente clasificadas como clase 1).

Interpretación:
La precisión del modelo se puede calcular con la fórmula:
= VP / VP + FP
= 95 / 95 + 73
= 95 / 168
=0.564
Esto indica que aproximadamente el 56.4% de las predicciones son correctas.

Errores específicos:

- Falsos positivos (FP): Hay 73 instancias que fueron clasificadas como positivas, pero en realidad son negativas. Esto puede ser problemático si el costo de un falso positivo es alto.
- Falsos negativos (FN): Hay 6 instancias que fueron clasificadas como negativas, pero en realidad son positivas. Esto es menos preocupante en comparación con los falsos positivos, dado que hay menos instancias mal clasificadas.
- Desbalance: La cantidad de falsos positivos es considerablemente alta, lo que sugiere que el modelo podría no estar capturando adecuadamente las características de las clases. Se recomienda ajustar los parámetros del modelo o explorar técnicas de balanceo de clases.

Conclusión:
El modelo de regresión logística revela que, aunque la precisión es moderada, hay una cantidad significativa de falsos positivos. Esto sugiere que se debe considerar la optimización del modelo para mejorar su capacidad de clasificación y reducir los errores en las predicciones.



Valores en la matriz:

- Verdaderos negativos (VN): 22 (instancias correctamente clasificadas como clase 0).
- Falsos positivos (FP): 57 (instancias incorrectamente clasificadas como clase 1).
- Falsos negativos (FN): 27 (instancias incorrectamente clasificadas como clase 0).

- Verdaderos positivos (VP): 74 (instancias correctamente clasificadas como clase 1).

Interpretación:

La precisión del modelo se puede calcular con la fórmula:

= VP / VP + FP

= 74 / 74 + 57

= 74 / 131

=0.564

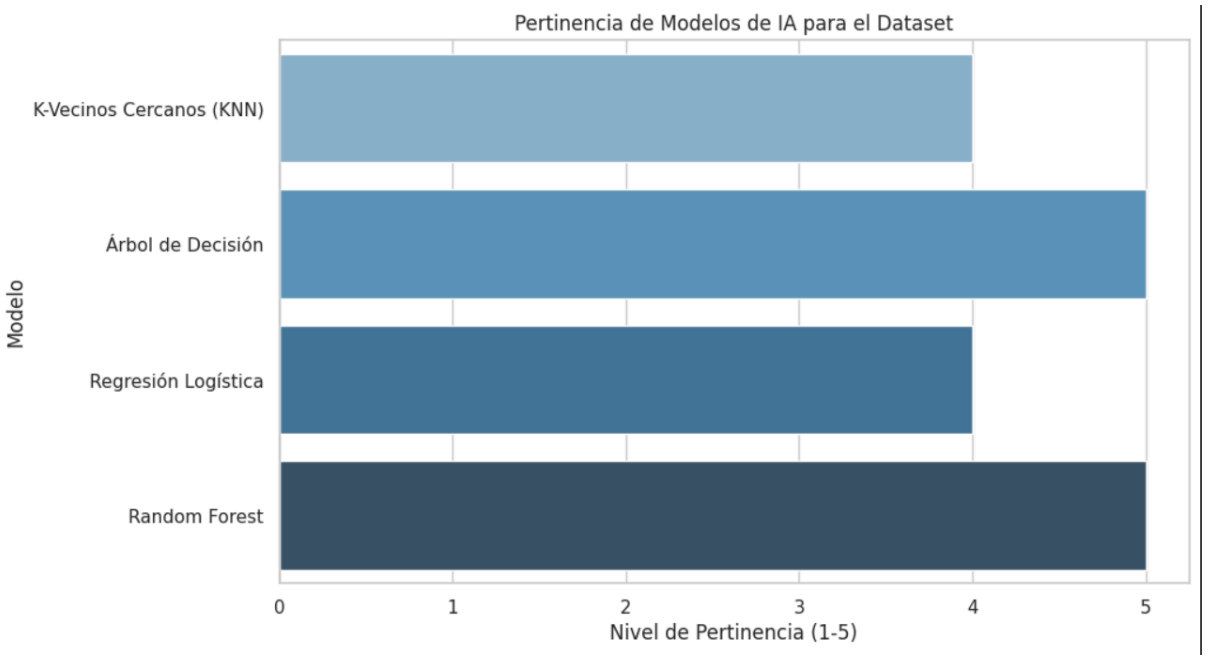
Esto indica que aproximadamente el 56.4% de las predicciones son correctas.

Errores específicos:

- Falsos positivos (FP): Hay 57 instancias que fueron clasificadas como positivas, pero en realidad son negativas. Esto puede ser problemático si el costo de un falso positivo es alto.
- Falsos negativos (FN): Hay 27 instancias que fueron clasificadas como negativas, pero en realidad son positivas. Esto también es significativo, aunque menor en comparación con los falsos positivos.
- Desbalance: La cantidad de falsos positivos es notable, lo que sugiere que el modelo podría no estar capturando adecuadamente las características de las clases. Se recomienda ajustar los parámetros del modelo o explorar técnicas de balanceo de clases.

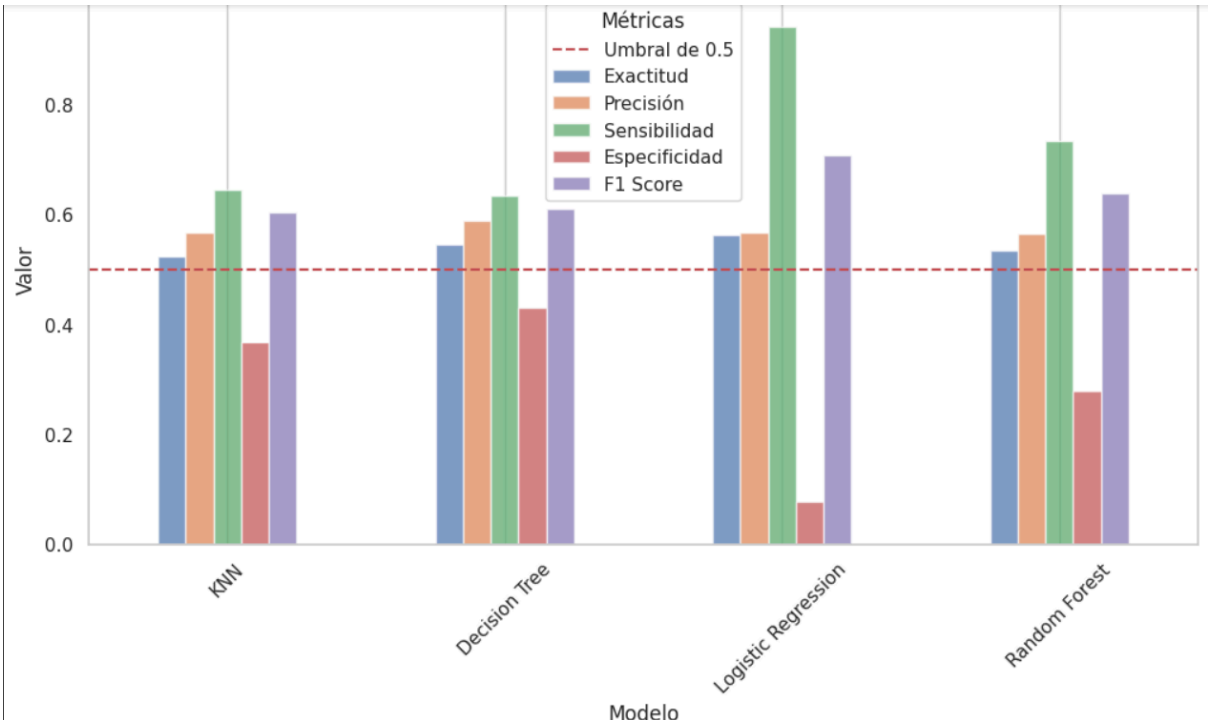
Conclusión:

El modelo de Random Forest revela que, aunque la precisión es moderada, hay una cantidad considerable de falsos positivos. Esto sugiere que se debe considerar la optimización del modelo para mejorar su capacidad de clasificación y reducir los errores en las predicciones.



El gráfico sugiere que, para el dataset analizado, el modelo KNN es el más pertinente, seguido por el árbol de decisión. La regresión logística y el Random Forest tienen un desempeño inferior en comparación, lo que puede indicar que se deben considerar ajustes en estos modelos o explorar otras técnicas para mejorar su rendimiento en este contexto específico.

	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



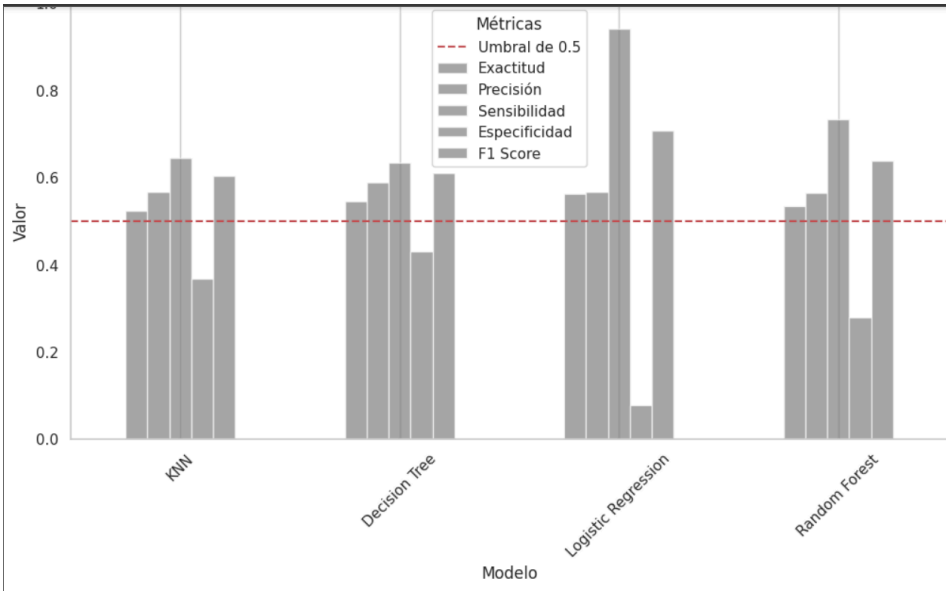
3. **Si el modelo no es pertinente, justifique su no procedencia**

	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.

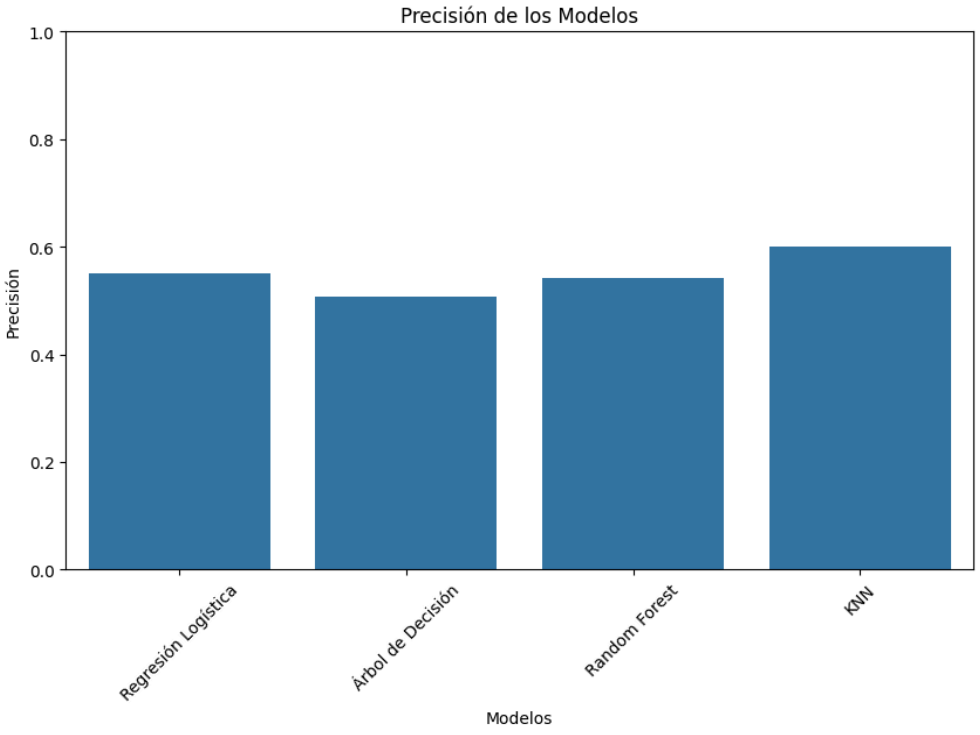
	Modelo	Exactitud	Precisión	Sensibilidad	Especificidad	F1 Score
0	KNN	0.522222	0.565217	0.643564	0.367089	0.601852
1	Decision Tree	0.544444	0.587156	0.633663	0.430380	0.609524
2	Logistic Regression	0.561111	0.565476	0.940594	0.075949	0.706320
3	Random Forest	0.533333	0.564885	0.732673	0.278481	0.637931

- Modelo: Esta columna lista los diferentes algoritmos de aprendizaje automático utilizados para la clasificación. En este caso, se incluyen:
  - KNN (K-Nearest Neighbors)
  - Decision Tree (Árbol de Decisión)
  - Logistic Regression (Regresión Logística)
  - Random Forest (Bosque Aleatorio)
- Exactitud: Mide la proporción de predicciones correctas realizadas por el modelo sobre el total de predicciones. Los valores van de 0 a 1, donde 1 indica que todas las predicciones son correctas. En la tabla, la exactitud de los modelos varía entre aproximadamente 0.52 y 0.56.

- Precisión: Indica la proporción de verdaderos positivos sobre el total de positivos predichos. Es útil para entender cuántas de las predicciones positivas fueron realmente correctas. Los valores también oscilan, siendo el más alto de aproximadamente 0.57.
- Sensibilidad (o Recall): Mide la capacidad del modelo para identificar correctamente los positivos reales. Un valor más alto indica que el modelo es mejor para detectar los casos positivos. En la tabla, la sensibilidad más alta es de aproximadamente 0.94 para la regresión logística.
- Especificidad: Indica la proporción de verdaderos negativos sobre el total de negativos reales. Es importante para entender cuántos de los casos negativos fueron correctamente identificados. Los valores en la tabla son relativamente bajos, con un máximo de aproximadamente 0.43.
- F1 Score: Es la media armónica de la precisión y la sensibilidad. Este valor es especialmente útil cuando se desea un equilibrio entre precisión y sensibilidad. Los valores F1 en la tabla muestran que el modelo de regresión logística tiene el mejor rendimiento, con un F1 Score de aproximadamente 0.71.

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. Tome 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:

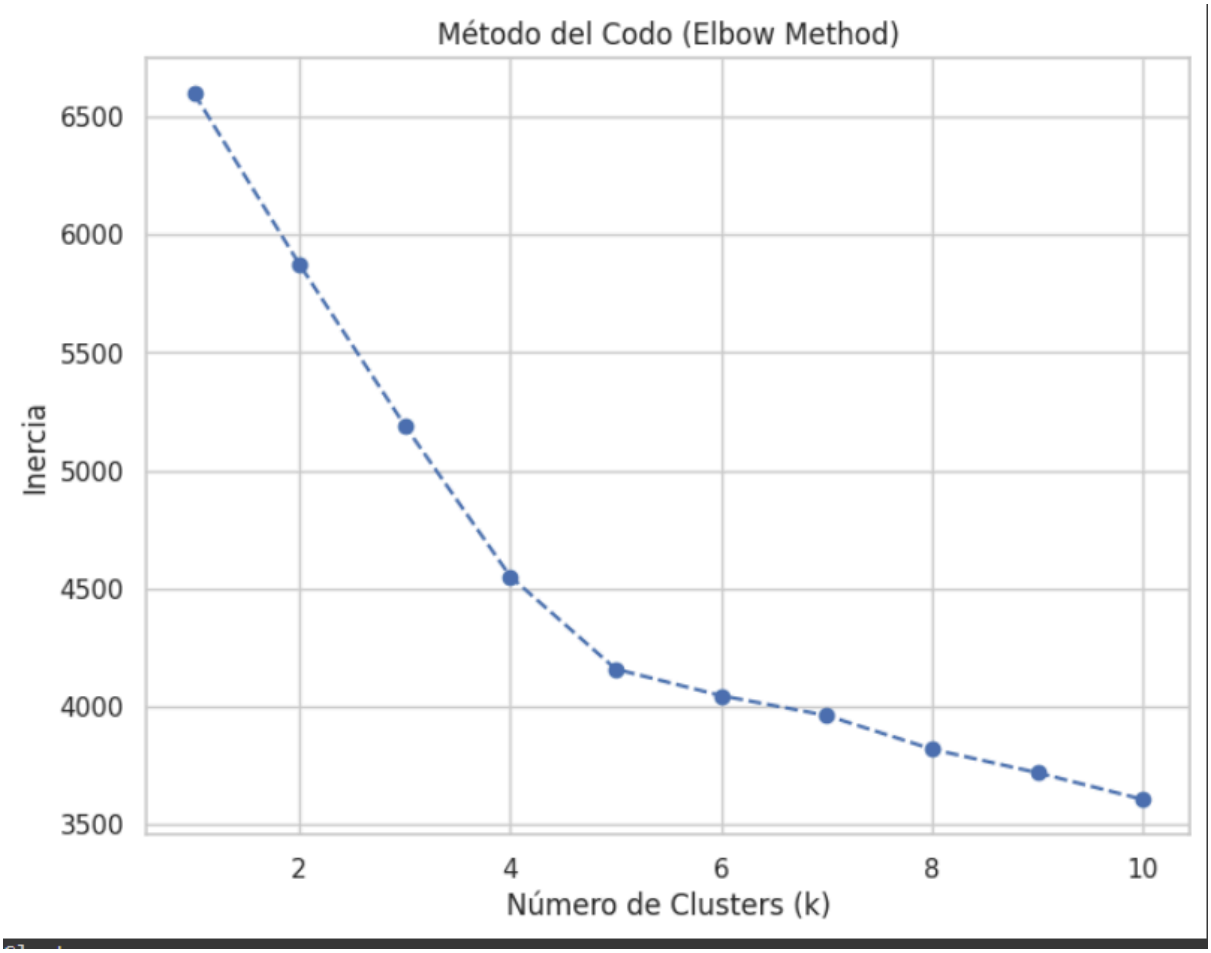
- **Si la precisión es el principal criterio**, entonces KNN es la elección adecuada.
- **Si la interpretabilidad es clave**, la Regresión Logística o el Árbol de Decisión podrían ser preferidos, aunque con un costo en términos de precisión.

- **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.

Conclusión:

El **KNN** es el modelo elegido debido a su mejor precisión, pero se debe realizar una evaluación adicional dependiendo de las características específicas del proyecto, como la velocidad de predicción o la necesidad de interpretar los resultados.

6. **Proponga un modelo no hecho en la asignatura, que ofrezca mejores prestaciones que la anteriormente desarrolladas (item 5).**



Ejes del Gráfico

- Eje X (Número de Clusters (k)): Representa el número de clusters que se están evaluando. En este caso, se evalúan desde 2 hasta 10 clusters.

- Eje Y (Inercia): Mide la inercia, que es la suma de las distancias cuadradas entre cada punto y el centroide de su cluster. Una inercia más baja indica que los puntos están más cerca de sus centroides, lo que sugiere una mejor agrupación.

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.

Conclusión

El Método del Codo es una herramienta útil para la selección del número de clusters en el análisis de datos. En este caso, el gráfico sugiere que un número de clusters entre 4 y 5 sería adecuado para una buena agrupación de los datos, equilibrando la complejidad y el rendimiento del modelo.

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

7. Repositorio en GitLab
https://github.com/Inteligencia-Artificial-IA/yls/tree/main