



INSTITUTO POLITÉCNICO NACIONAL

ESCUELA SUPERIOR DE CÓMPUTO
INGENIERIA EN SISTEMAS COMPUTACIONALES

Inteligencia Artificial

Ejercicio de laboratorio 8
“Comparación de Clasificadores”

Presentan

Gómez Hernández, Alan Javier
Hernández Pérez, Juan Manuel
Jiménez Cruz, Daniel
Rivas Carrera, Diana Laura

Docente

Andrés García Floriano



noviembre de 2023

Contents

Introducción	3
Desarrollo	4
I) Validación Holdout	4
A) Iris	4
B) Wine	5
C) Breast Cancer	6
II) Validación Cruzada	7
A) Iris	7
B) Wine	8
C) Breast Cancer	9
Conclusión.....	10

Introducción

1-NN (Vecino más cercano)

El clasificador 1-NN, también conocido como clasificador de 1 vecino más cercano, es un algoritmo de aprendizaje supervisado utilizado para problemas de clasificación.

Vecino más Cercano: Para clasificar un nuevo punto de datos, el algoritmo encuentra el punto más cercano en el conjunto de entrenamiento. Este punto más cercano es el "vecino más cercano."

Etiqueta del Vecino: La etiqueta (clase) asignada al nuevo punto es la misma que la etiqueta del vecino más cercano. En otras palabras, se asigna la clase del punto más cercano al nuevo punto

Clasificador K-NN

El clasificador K-NN, o clasificador de K vecinos más cercanos, es un algoritmo de aprendizaje supervisado utilizado para problemas de clasificación y también puede aplicarse a problemas de regresión.

Vecinos más Cercanos: En lugar de depender de un solo vecino cercano (como en el 1-NN), el clasificador K-NN considera los K vecinos más cercanos al nuevo punto de datos que se está clasificando.

Votación: Las etiquetas de los K vecinos más cercanos se toman en cuenta, y la clase más frecuente entre esos vecinos se asigna como la clase del nuevo punto.

Clasificador Bayesiano (Naive Bayes)

El clasificador bayesiano, también conocido como Naive Bayes, es un tipo de algoritmo de aprendizaje supervisado basado en el teorema de Bayes. Es particularmente efectivo para problemas de clasificación y se utiliza comúnmente en tareas como filtrado de spam y categorización de documentos

Teorema de Bayes: El clasificador Naive Bayes se basa en el teorema de Bayes, que describe la probabilidad condicional de un evento dado otro evento. En el contexto del aprendizaje automático, calcula la probabilidad de que una instancia pertenezca a una clase dada sus características.

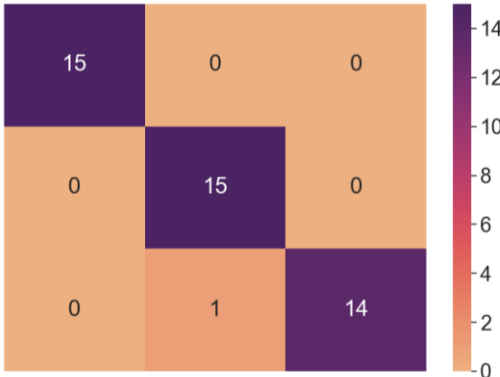
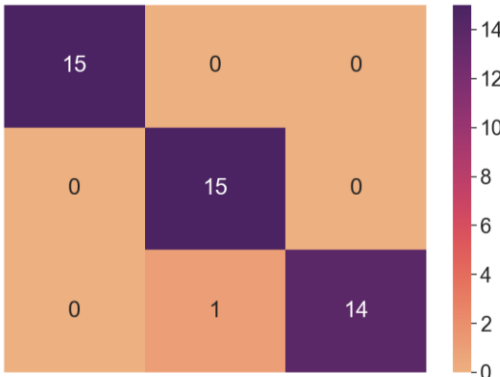
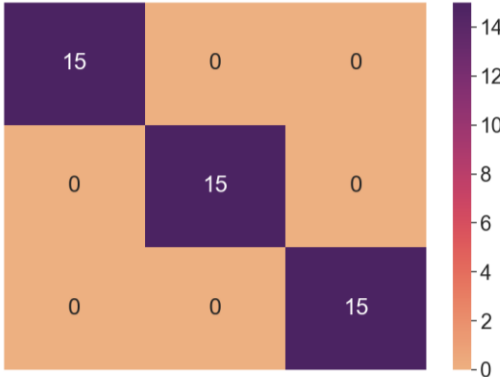
Suposición "Naive": El término "naive" (ingenuo) proviene de la suposición de independencia condicional entre todas las características dadas las clases. Esto significa que se asume que la presencia o ausencia de una característica no está relacionada con la presencia o ausencia de otras características, lo que simplifica el cálculo de las probabilidades.

Desarrollo

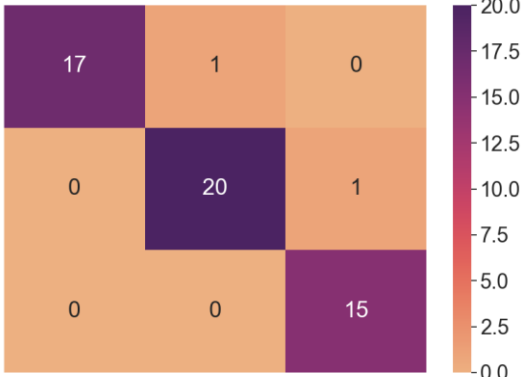
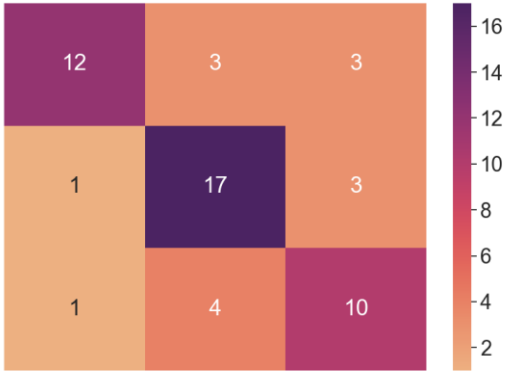

I) Validación Holdout

En esta primera parte se dividió el conjunto de datos con una proporción de 70% para el conjunto de entrenamiento y 30% para el conjunto de validación.

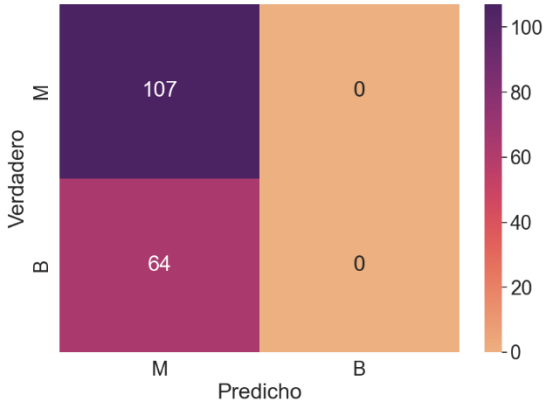
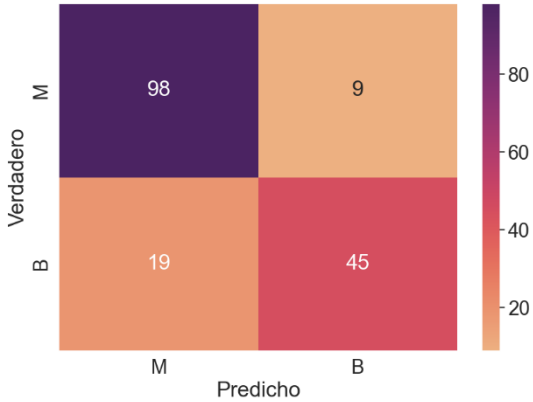
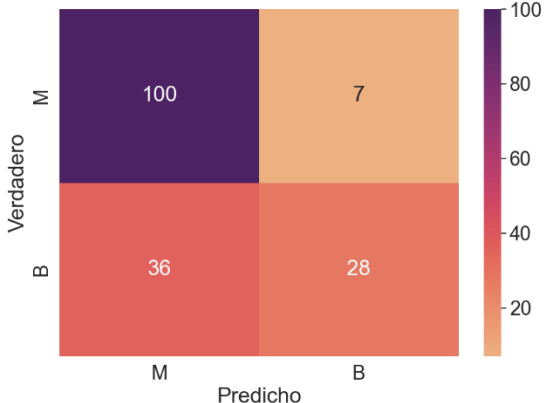
A) Iris

Clasificador	Acc.	Matriz de Confusión																
Naive-Bayes	97.78%	<div><div><div>setosa</div><div>Verdadero</div><div>versicolor</div><div>virginica</div></div><div><div><div>setosa</div><div>versicolor</div><div>virginica</div></div><div>Predicho</div></div><table><tr><th>Verdadero \ Predicho</th><th>setosa</th><th>versicolor</th><th>virginica</th></tr><tr><th>setosa</th><td>15</td><td>0</td><td>0</td></tr><tr><th>versicolor</th><td>0</td><td>15</td><td>0</td></tr><tr><th>virginica</th><td>0</td><td>1</td><td>14</td></tr></table></div>	Verdadero \ Predicho	setosa	versicolor	virginica	setosa	15	0	0	versicolor	0	15	0	virginica	0	1	14
Verdadero \ Predicho	setosa	versicolor	virginica															
setosa	15	0	0															
versicolor	0	15	0															
virginica	0	1	14															
1-NN	97.78%	<div><div><div>setosa</div><div>Verdadero</div><div>versicolor</div><div>virginica</div></div><div><div><div>setosa</div><div>versicolor</div><div>virginica</div></div><div>Predicho</div></div><table><tr><th>Verdadero \ Predicho</th><th>setosa</th><th>versicolor</th><th>virginica</th></tr><tr><th>setosa</th><td>15</td><td>0</td><td>0</td></tr><tr><th>versicolor</th><td>0</td><td>15</td><td>0</td></tr><tr><th>virginica</th><td>0</td><td>1</td><td>14</td></tr></table></div>	Verdadero \ Predicho	setosa	versicolor	virginica	setosa	15	0	0	versicolor	0	15	0	virginica	0	1	14
Verdadero \ Predicho	setosa	versicolor	virginica															
setosa	15	0	0															
versicolor	0	15	0															
virginica	0	1	14															
5-NN	100%	<div><div><div>setosa</div><div>Verdadero</div><div>versicolor</div><div>virginica</div></div><div><div><div>setosa</div><div>versicolor</div><div>virginica</div></div><div>Predicho</div></div><table><tr><th>Verdadero \ Predicho</th><th>setosa</th><th>versicolor</th><th>virginica</th></tr><tr><th>setosa</th><td>15</td><td>0</td><td>0</td></tr><tr><th>versicolor</th><td>0</td><td>15</td><td>0</td></tr><tr><th>virginica</th><td>0</td><td>0</td><td>15</td></tr></table></div>	Verdadero \ Predicho	setosa	versicolor	virginica	setosa	15	0	0	versicolor	0	15	0	virginica	0	0	15
Verdadero \ Predicho	setosa	versicolor	virginica															
setosa	15	0	0															
versicolor	0	15	0															
virginica	0	0	15															

B) Wine

Clasificador	Acc.	Matriz de Confusión																
Naive-Bayes	96.30%	<div><div><div>Verdadero</div><div><div><div>1</div><div>2</div><div>3</div></div><div><div><div>1</div><div>2</div><div>3</div></div><div>Predicho</div></div></div><div><table><tr><th></th><th>1</th><th>2</th><th>3</th></tr><tr><th>1</th><td>17</td><td>1</td><td>0</td></tr><tr><th>2</th><td>0</td><td>20</td><td>1</td></tr><tr><th>3</th><td>0</td><td>0</td><td>15</td></tr></table></div></div></div>		1	2	3	1	17	1	0	2	0	20	1	3	0	0	15
	1	2	3															
1	17	1	0															
2	0	20	1															
3	0	0	15															
1-NN	72.22%	<div><div><div>Verdadero</div><div><div><div>1</div><div>2</div><div>3</div></div><div><div><div>1</div><div>2</div><div>3</div></div><div>Predicho</div></div></div><div><table><tr><th></th><th>1</th><th>2</th><th>3</th></tr><tr><th>1</th><td>12</td><td>3</td><td>3</td></tr><tr><th>2</th><td>1</td><td>17</td><td>3</td></tr><tr><th>3</th><td>1</td><td>4</td><td>10</td></tr></table></div></div></div>		1	2	3	1	12	3	3	2	1	17	3	3	1	4	10
	1	2	3															
1	12	3	3															
2	1	17	3															
3	1	4	10															
5-NN	72.22%	<div><div><div>Verdadero</div><div><div><div>1</div><div>2</div><div>3</div></div><div><div><div>1</div><div>2</div><div>3</div></div><div>Predicho</div></div></div><div><table><tr><th></th><th>1</th><th>2</th><th>3</th></tr><tr><th>1</th><td>18</td><td>0</td><td>0</td></tr><tr><th>2</th><td>2</td><td>13</td><td>6</td></tr><tr><th>3</th><td>4</td><td>3</td><td>8</td></tr></table></div></div></div>		1	2	3	1	18	0	0	2	2	13	6	3	4	3	8
	1	2	3															
1	18	0	0															
2	2	13	6															
3	4	3	8															

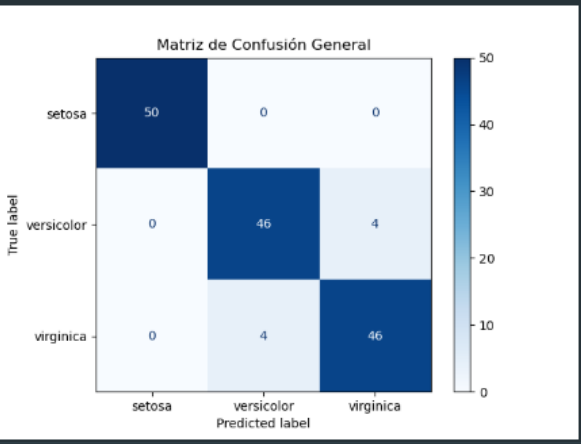
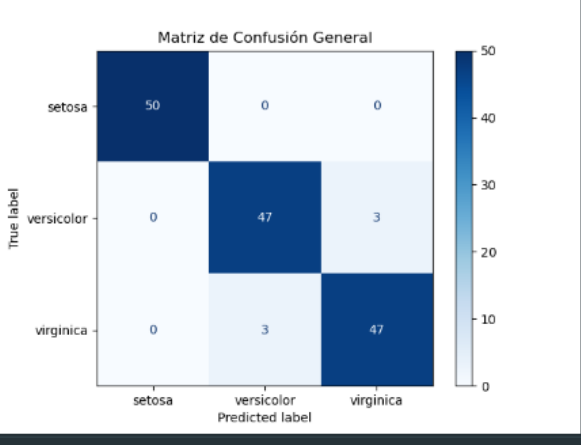
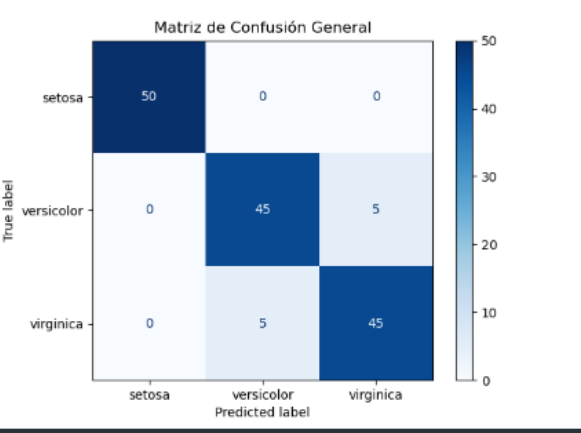
C) Breast Cancer

Clasificador	Acc.	Matriz de Confusión
Naive-Bayes	62.57%	 <p>Verdadero</p> <p>M B</p> <p>M B Predicho</p>
1-NN	83.63%	 <p>Verdadero</p> <p>M B</p> <p>M B Predicho</p>
7-NN	74.85%	 <p>Verdadero</p> <p>M B</p> <p>M B Predicho</p>

II) Validación Cruzada

Para esta segunda parte se dividió el conjunto de datos en 10 pliegues para realizar la validación cruzada.

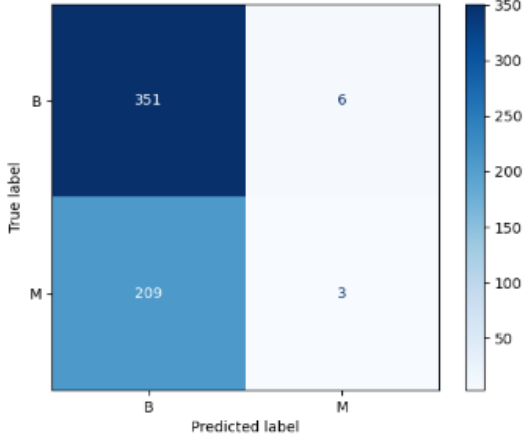
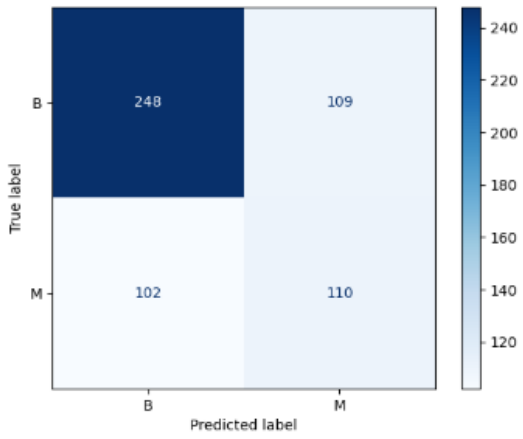
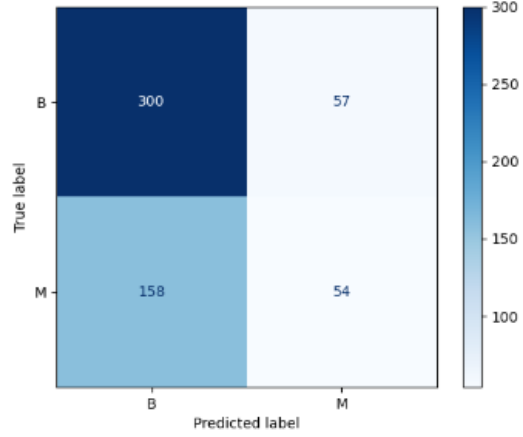
A) Iris

Clasificador	Acc.	Matriz de Confusión																	
Naive-Bayes	94%		 <p>Matriz de Confusión General</p> <table border="1"> <thead> <tr> <th>True label \ Predicted label</th><th>setosa</th><th>versicolor</th><th>virginica</th></tr> </thead> <tbody> <tr> <th>setosa</th><td>50</td><td>0</td><td>0</td></tr> <tr> <th>versicolor</th><td>0</td><td>46</td><td>4</td></tr> <tr> <th>virginica</th><td>0</td><td>4</td><td>46</td></tr> </tbody> </table>	True label \ Predicted label	setosa	versicolor	virginica	setosa	50	0	0	versicolor	0	46	4	virginica	0	4	46
True label \ Predicted label	setosa	versicolor	virginica																
setosa	50	0	0																
versicolor	0	46	4																
virginica	0	4	46																
1-NN	96%		 <p>Matriz de Confusión General</p> <table border="1"> <thead> <tr> <th>True label \ Predicted label</th><th>setosa</th><th>versicolor</th><th>virginica</th></tr> </thead> <tbody> <tr> <th>setosa</th><td>50</td><td>0</td><td>0</td></tr> <tr> <th>versicolor</th><td>0</td><td>47</td><td>3</td></tr> <tr> <th>virginica</th><td>0</td><td>3</td><td>47</td></tr> </tbody> </table>	True label \ Predicted label	setosa	versicolor	virginica	setosa	50	0	0	versicolor	0	47	3	virginica	0	3	47
True label \ Predicted label	setosa	versicolor	virginica																
setosa	50	0	0																
versicolor	0	47	3																
virginica	0	3	47																
5-NN	93%		 <p>Matriz de Confusión General</p> <table border="1"> <thead> <tr> <th>True label \ Predicted label</th><th>setosa</th><th>versicolor</th><th>virginica</th></tr> </thead> <tbody> <tr> <th>setosa</th><td>50</td><td>0</td><td>0</td></tr> <tr> <th>versicolor</th><td>0</td><td>45</td><td>5</td></tr> <tr> <th>virginica</th><td>0</td><td>5</td><td>45</td></tr> </tbody> </table>	True label \ Predicted label	setosa	versicolor	virginica	setosa	50	0	0	versicolor	0	45	5	virginica	0	5	45
True label \ Predicted label	setosa	versicolor	virginica																
setosa	50	0	0																
versicolor	0	45	5																
virginica	0	5	45																

B) Wine

Clasificador	Acc.	Matriz de Confusión																	
Naive-Bayes	96.1%		<p>Matriz de Confusión General</p> <table border="1"> <thead> <tr> <th>True label \ Predicted label</th><th>1</th><th>2</th><th>3</th></tr> </thead> <tbody> <tr> <th>1</th><td>56</td><td>3</td><td>0</td></tr> <tr> <th>2</th><td>1</td><td>67</td><td>3</td></tr> <tr> <th>3</th><td>0</td><td>0</td><td>48</td></tr> </tbody> </table>	True label \ Predicted label	1	2	3	1	56	3	0	2	1	67	3	3	0	0	48
True label \ Predicted label	1	2	3																
1	56	3	0																
2	1	67	3																
3	0	0	48																
1-NN	67.67%		<p>Matriz de Confusión General</p> <table border="1"> <thead> <tr> <th>True label \ Predicted label</th><th>1</th><th>2</th><th>3</th></tr> </thead> <tbody> <tr> <th>1</th><td>48</td><td>7</td><td>4</td></tr> <tr> <th>2</th><td>5</td><td>52</td><td>14</td></tr> <tr> <th>3</th><td>7</td><td>20</td><td>21</td></tr> </tbody> </table>	True label \ Predicted label	1	2	3	1	48	7	4	2	5	52	14	3	7	20	21
True label \ Predicted label	1	2	3																
1	48	7	4																
2	5	52	14																
3	7	20	21																
5-NN	63.11%		<p>Matriz de Confusión General</p> <table border="1"> <thead> <tr> <th>True label \ Predicted label</th><th>1</th><th>2</th><th>3</th></tr> </thead> <tbody> <tr> <th>1</th><td>51</td><td>2</td><td>6</td></tr> <tr> <th>2</th><td>6</td><td>46</td><td>19</td></tr> <tr> <th>3</th><td>8</td><td>24</td><td>16</td></tr> </tbody> </table>	True label \ Predicted label	1	2	3	1	51	2	6	2	6	46	19	3	8	24	16
True label \ Predicted label	1	2	3																
1	51	2	6																
2	6	46	19																
3	8	24	16																

C) Breast Cancer

Clasificador	Acc.	Matriz de Confusión										
Naive-Bayes	62.23%		<div><p>Matriz de Confusión General</p><table><tr><th></th><th>B</th><th>M</th></tr><tr><th>B</th><td>351</td><td>6</td></tr><tr><th>M</th><td>209</td><td>3</td></tr></table></div>		B	M	B	351	6	M	209	3
	B	M										
B	351	6										
M	209	3										
1-NN	62.86%		<div><p>Matriz de Confusión General</p><table><tr><th></th><th>B</th><th>M</th></tr><tr><th>B</th><td>248</td><td>109</td></tr><tr><th>M</th><td>102</td><td>110</td></tr></table></div>		B	M	B	248	109	M	102	110
	B	M										
B	248	109										
M	102	110										
5-NN	62.25%		<div><p>Matriz de Confusión General</p><table><tr><th></th><th>B</th><th>M</th></tr><tr><th>B</th><td>300</td><td>57</td></tr><tr><th>M</th><td>158</td><td>54</td></tr></table></div>		B	M	B	300	57	M	158	54
	B	M										
B	300	57										
M	158	54										

Conclusión

Al analizar los resultados obtenidos de la comparación entre los clasificadores Naive Bayes (NB) y K-Vecinos más Cercanos (K-NN) con distintos números de vecinos en los conjuntos de datos de Iris, Wine y Breast Cancer, se observan varias tendencias interesantes y lecciones importantes.

En el caso del conjunto de datos de Iris, tanto NB como 1-NN presentan un alto rendimiento, con una ligera ventaja para el 1-NN en la validación Holdout. Sin embargo, en la validación cruzada, el 1-NN supera ligeramente a NB, aunque ambos superan a 5-NN. Este patrón sugiere que mientras que NB es un modelo robusto, el ajuste fino de los vecinos en K-NN puede ofrecer mejoras significativas en algunos casos.

Por otro lado, en el dataset de Wine, NB muestra una consistente superioridad sobre K-NN, tanto con 1 como con 5 vecinos, en ambas validaciones Holdout y cruzada. Esto indica que el supuesto de independencia de características en NB puede ser menos restrictivo en este caso particular en comparación con la dependencia de la selección de vecinos en K-NN.

En cuanto al dataset de Breast Cancer, se observa que mientras NB muestra un rendimiento moderado, 1-NN se destaca significativamente en la validación Holdout. No obstante, en la validación cruzada, los tres modelos presentan resultados similares y relativamente bajos, lo que podría indicar una mayor variabilidad en los datos que afecta el rendimiento del modelo.

La elección de la validación Holdout con una proporción de 70/30, especialmente en datasets pequeños, tiene sus propias implicaciones. Existe un riesgo considerable de overfitting, ya que el modelo podría ajustarse excesivamente a un conjunto de entrenamiento más grande y no generalizar bien en un conjunto de prueba más pequeño. Asimismo, el underfitting también puede ser una preocupación si el conjunto de entrenamiento no captura suficientemente la variabilidad o las características representativas de todas las clases.

Por otro lado, la validación cruzada, especialmente con 10 pliegues, ofrece una evaluación más robusta y menos sesgada del desempeño del modelo. Al utilizar cada instancia de los datos tanto para entrenamiento como para prueba, esta metodología es particularmente valiosa en datasets pequeños, donde cada punto de datos es crucial para entender la variabilidad general.

No existe un clasificador universalmente mejor; la efectividad depende del conjunto de datos específico y sus características. La elección del método de validación tiene un impacto significativo en la evaluación del rendimiento del modelo. La validación cruzada suele ser preferible en conjuntos de datos pequeños para reducir los riesgos de overfitting y underfitting. Además, es esencial considerar las características y supuestos del modelo al seleccionar un clasificador para un problema específico, como la independencia de características en NB o la elección del número de vecinos en K-NN.