



INSTITUTO POLITÉCNICO NACIONAL ESCUELA SUPERIOR DE CÓMPUTO

INGENIERIA EN SISTEMAS COMPUTACIONALES

Ejercicio de Laboratorio No. 9

Presentan

GÓMEZ HERNÁNDEZ ALAN JAVIER

HERNÁNDEZ PÉREZ JUAN MANUEL

JIMÉNEZ CRUZ DANIEL

RIVAS CARRERA DIANA LAURA

Profesor

ANDRÉS GARCÍA FLORIANO



Noviembre 2023

Introducción

La aplicación de algoritmos de aprendizaje automático para el análisis de datos se ha convertido en una herramienta fundamental en diversos campos, incluyendo la biología, la medicina y la economía. Los modelos seleccionados para práctica tarea incluyen el Árbol de clasificación ID5 o J48, Random Forest, y tres variantes del SVM (Support Vector Machine) con diferentes kernels: lineal, Gaussiano, y polinomial.

Cada uno de estos modelos posee características únicas y es adecuado para ciertos tipos de problemas. Por ejemplo, los **árboles de decisión** son útiles para entender las decisiones del modelo, mientras que los **SVM** son eficientes en espacios de alta dimensión. **Random Forest**, siendo un ensamble de árboles de decisión, es conocido por su robustez y su capacidad para evitar el sobreajuste.

La selección de parámetros óptimos para cada modelo es crucial para lograr el mejor rendimiento posible. Esto incluye, pero no se limita a, la elección del kernel en SVM y la cantidad de árboles en Random Forest. Los conjuntos de datos seleccionados (Iris, Wine, y Breast Cancer) son ampliamente utilizados como benchmarks en el campo del aprendizaje automático, proporcionando una base sólida para la comparación y evaluación de modelos.

Para validar los modelos y estimar su rendimiento en datos no vistos, se utilizarán dos métodos de validación comunes: Hold-Out con una división 70/30 y 10-Fold Cross Validation. Estos métodos proporcionan formas diferentes de entender cómo los modelos generalizarán a nuevos datos.

Las métricas de rendimiento incluirán la precisión (Accuracy), medida como un porcentaje o en una escala de 0 a 1, y la matriz de confusión, que proporciona una visión detallada de la clasificación en términos de verdaderos positivos, falsos positivos, verdaderos negativos y falsos negativos.

Árbol de Clasificación ID5 o J48

Los árboles de clasificación son modelos de aprendizaje supervisado que dividen el conjunto de datos en subconjuntos más pequeños basándose en las características descriptivas. Se representan en forma de árboles, donde cada nodo interno representa una decisión basada en una característica, y cada hoja representa una etiqueta de clase.

Los algoritmos ID5/J48 emplean un enfoque de entropía o ganancia de información para seleccionar la característica que mejor divide el conjunto de datos en cada paso.

Random Forest

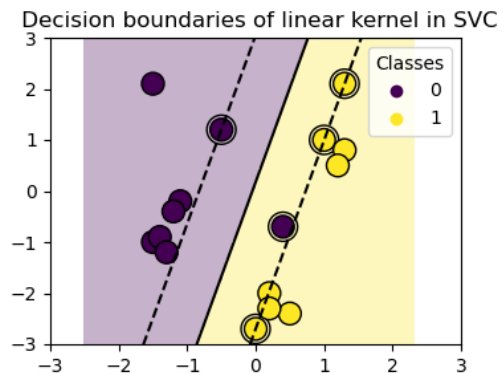
Random Forest es un algoritmo de ensamble que construye múltiples árboles de decisión durante el entrenamiento y produce la clase que es el modo de las clases (clasificación) o la media de las predicciones (regresión) de los árboles individuales.

Mejora la precisión del modelo y controla el sobreajuste mediante la combinación de varios árboles de decisión. La diversidad entre los árboles se logra mediante el uso de subconjuntos aleatorios de características y ejemplos.

SVM con Kernel Lineal

Las Máquinas de Soporte Vectorial (SVM) son modelos de aprendizaje supervisado utilizados para la clasificación y regresión. En su forma básica, tratan de encontrar el hiperplano que mejor separa las clases en el espacio de características.

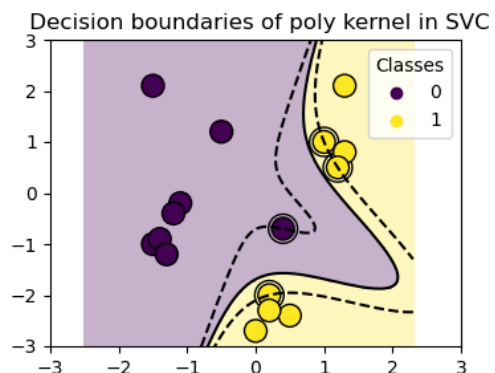
Utiliza un kernel lineal que es efectivamente el producto escalar en el espacio de entrada. Es adecuado para datos que son linealmente separables.



SVM con Kernel Gaussiano

También conocido como el kernel radial de base (RBF), transforma el espacio de características en una dimensión más alta donde es más probable que los datos sean linealmente separables.

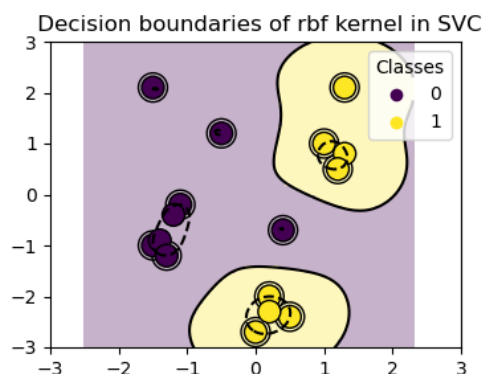
Utiliza una función de similitud que mide la distancia entre pares de muestras en un espacio caracterizado por una función gaussiana.



SVM con Kernel Polinomial

Este kernel representa las interacciones de todas las características hasta un cierto grado del polinomio. Permite modelar no solo las características individuales sino también las combinaciones de ellas.

El grado del polinomio determina la complejidad del modelo. Un grado más alto puede capturar relaciones más complejas, pero también aumenta el riesgo de sobreajuste.





Desarrollo

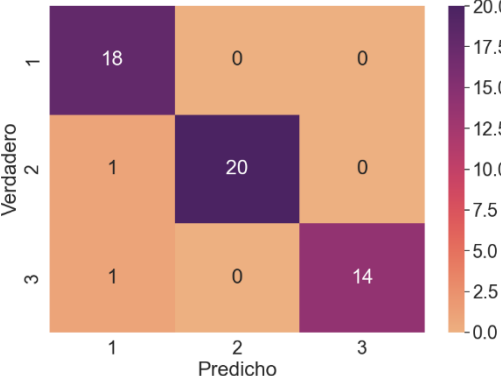
A) Holdout

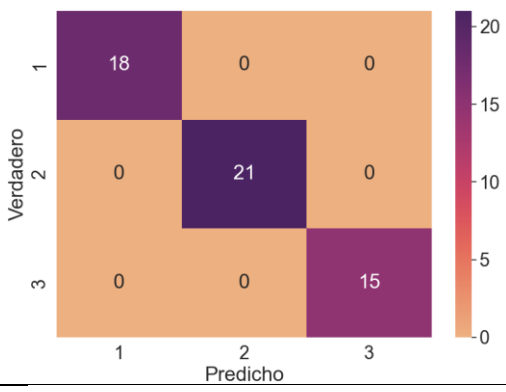
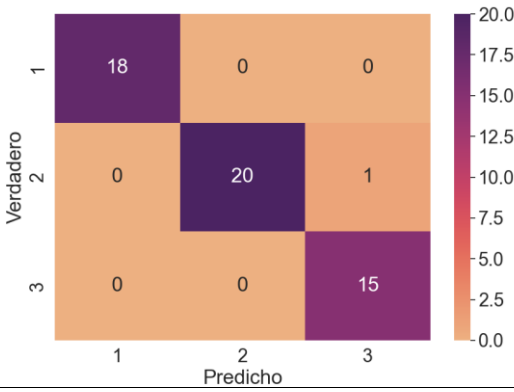
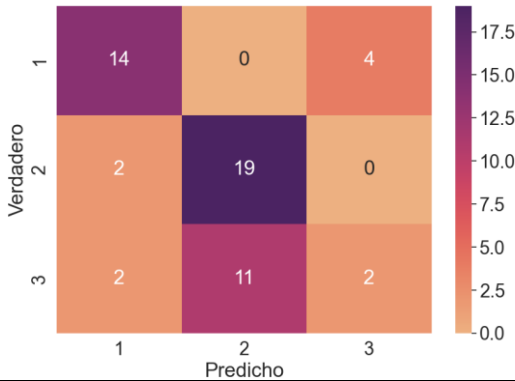
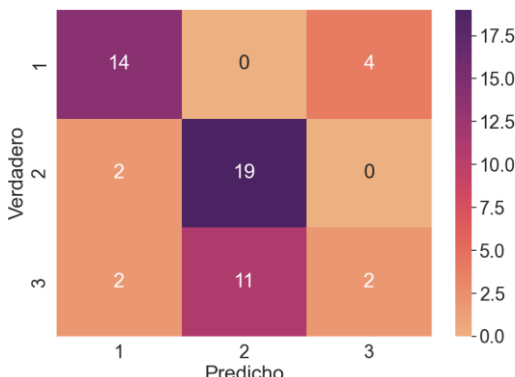
1) Iris

Clasificador	Accuracy	Matriz de Confusión																
Árbol de Clasificación J48	95.56%	<div><div><div>Verdadero</div><div>setosa</div><div>versicolor</div><div>virginica</div></div><div><div>setosa</div><div>versicolor</div><div>virginica</div></div><div>Predicho</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>2</td><td>13</td></tr></table></div>	Verdadero \ Predicho	setosa	versicolor	virginica	setosa	15	0	0	versicolor	0	15	0	virginica	0	2	13
Verdadero \ Predicho	setosa	versicolor	virginica															
setosa	15	0	0															
versicolor	0	15	0															
virginica	0	2	13															
Random Forest	97.78%	<div><div><div>Verdadero</div><div>setosa</div><div>versicolor</div><div>virginica</div></div><div><div>setosa</div><div>versicolor</div><div>virginica</div></div><div>Predicho</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															
SVM (Kernel Lineal)	100%	<div><div><div>Verdadero</div><div>setosa</div><div>versicolor</div><div>virginica</div></div><div><div>setosa</div><div>versicolor</div><div>virginica</div></div><div>Predicho</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															

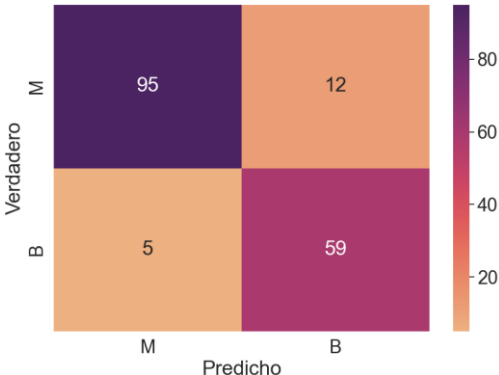
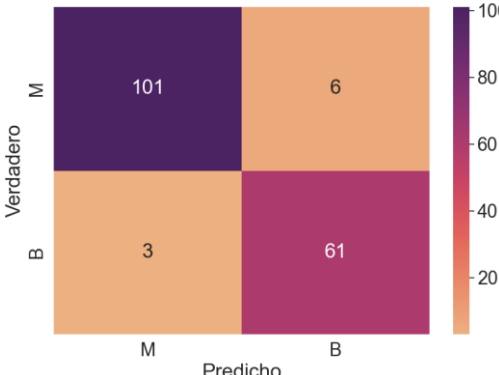
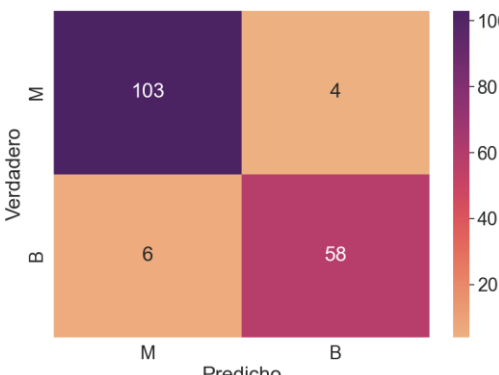
SVM (Kernel Gaussiano)	100%	
SVM (Kernel Polinomial) Grado: 2	100%	

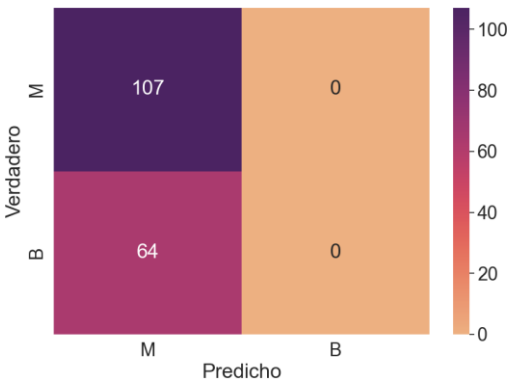
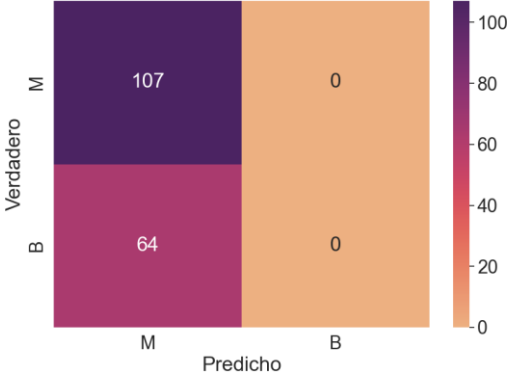
2) Wine

Clasificador	Accuracy	Matriz de Confusión
Árbol de Clasificación J48	96.44%	

Random Forest	100%	 <table><tr><th>Verdadero \ Predicho</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>0</td><td>21</td><td>0</td></tr><tr><th>3</th><td>0</td><td>0</td><td>15</td></tr></table>	Verdadero \ Predicho	1	2	3	1	18	0	0	2	0	21	0	3	0	0	15
Verdadero \ Predicho	1	2	3															
1	18	0	0															
2	0	21	0															
3	0	0	15															
SVM (Kernel Lineal)	98.15%	 <table><tr><th>Verdadero \ Predicho</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>0</td><td>20</td><td>1</td></tr><tr><th>3</th><td>0</td><td>0</td><td>15</td></tr></table>	Verdadero \ Predicho	1	2	3	1	18	0	0	2	0	20	1	3	0	0	15
Verdadero \ Predicho	1	2	3															
1	18	0	0															
2	0	20	1															
3	0	0	15															
SVM (Kernel Gaussiano)	64.81%	 <table><tr><th>Verdadero \ Predicho</th><th>1</th><th>2</th><th>3</th></tr><tr><th>1</th><td>14</td><td>0</td><td>4</td></tr><tr><th>2</th><td>2</td><td>19</td><td>0</td></tr><tr><th>3</th><td>2</td><td>11</td><td>2</td></tr></table>	Verdadero \ Predicho	1	2	3	1	14	0	4	2	2	19	0	3	2	11	2
Verdadero \ Predicho	1	2	3															
1	14	0	4															
2	2	19	0															
3	2	11	2															
SVM (Kernel Polinomial) Grado: 2	64.81%	 <table><tr><th>Verdadero \ Predicho</th><th>1</th><th>2</th><th>3</th></tr><tr><th>1</th><td>14</td><td>0</td><td>4</td></tr><tr><th>2</th><td>2</td><td>19</td><td>0</td></tr><tr><th>3</th><td>2</td><td>11</td><td>2</td></tr></table>	Verdadero \ Predicho	1	2	3	1	14	0	4	2	2	19	0	3	2	11	2
Verdadero \ Predicho	1	2	3															
1	14	0	4															
2	2	19	0															
3	2	11	2															

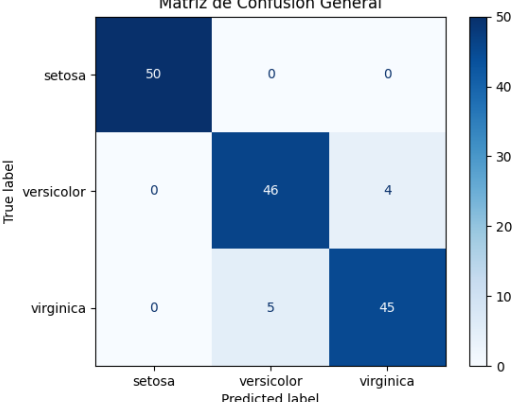
3) Breast Cancer

Clasificador	Accuracy	Matriz de Confusión									
Árbol de Clasificación J48	90.06%	 <table border="1"> <thead> <tr> <th></th> <th>Predicho M</th> <th>Predicho B</th> </tr> </thead> <tbody> <tr> <th>Verdadero M</th> <td>95</td> <td>12</td> </tr> <tr> <th>Verdadero B</th> <td>5</td> <td>59</td> </tr> </tbody> </table>		Predicho M	Predicho B	Verdadero M	95	12	Verdadero B	5	59
	Predicho M	Predicho B									
Verdadero M	95	12									
Verdadero B	5	59									
Random Forest	94.74%	 <table border="1"> <thead> <tr> <th></th> <th>Predicho M</th> <th>Predicho B</th> </tr> </thead> <tbody> <tr> <th>Verdadero M</th> <td>101</td> <td>6</td> </tr> <tr> <th>Verdadero B</th> <td>3</td> <td>61</td> </tr> </tbody> </table>		Predicho M	Predicho B	Verdadero M	101	6	Verdadero B	3	61
	Predicho M	Predicho B									
Verdadero M	101	6									
Verdadero B	3	61									
SVM (Kernel Lineal)	94.15%	 <table border="1"> <thead> <tr> <th></th> <th>Predicho M</th> <th>Predicho B</th> </tr> </thead> <tbody> <tr> <th>Verdadero M</th> <td>103</td> <td>4</td> </tr> <tr> <th>Verdadero B</th> <td>6</td> <td>58</td> </tr> </tbody> </table>		Predicho M	Predicho B	Verdadero M	103	4	Verdadero B	6	58
	Predicho M	Predicho B									
Verdadero M	103	4									
Verdadero B	6	58									

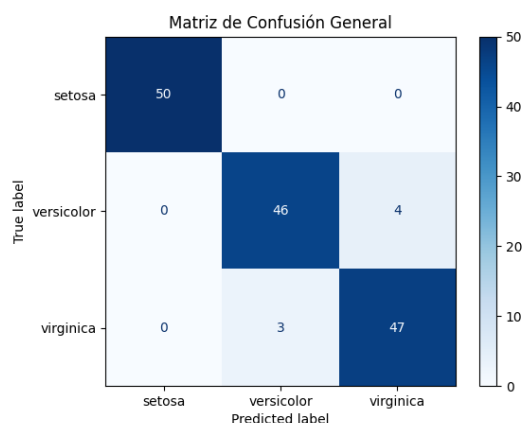
SVM (Kernel Gaussiano)	62.57%	 <p>Verdadero</p> <p>M B</p> <p>Predicho M B</p>
SVM (Kernel Polinomial) Grado: 2	62.57%	 <p>Verdadero</p> <p>M B</p> <p>Predicho M B</p>

B) Validación Cruzada

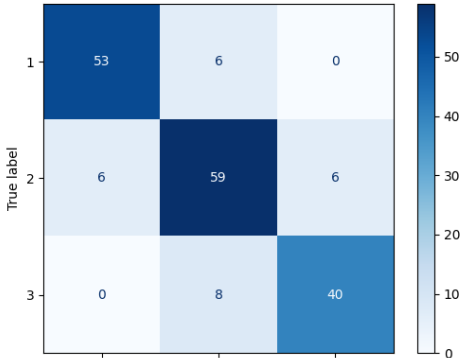
1) Iris

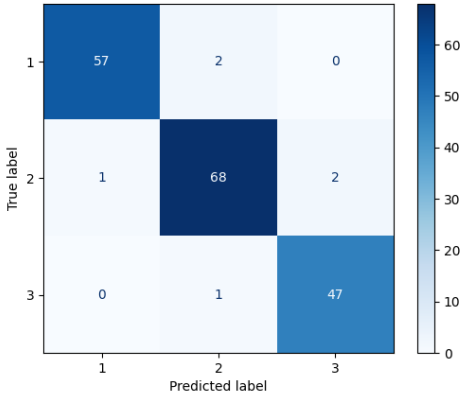
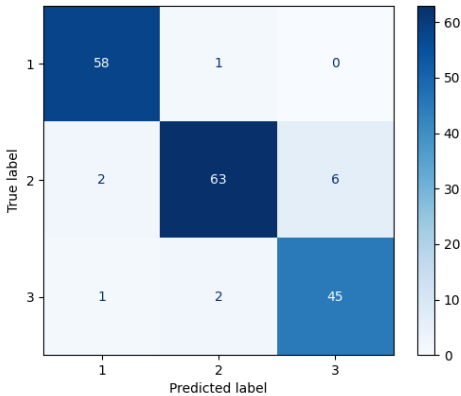
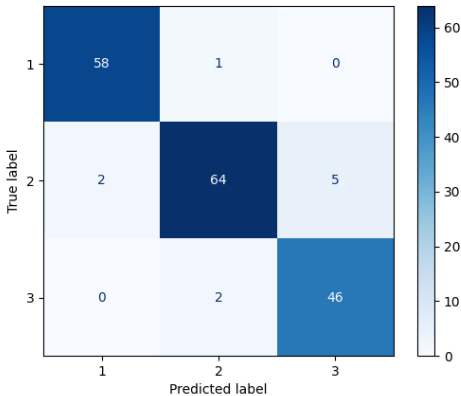
Clasificador	Accuracy	Matriz de Confusión
Árbol de Clasificación J48	94.00%	<p>Matriz de Confusión General</p>  <p>True label</p> <p>setosa versicolor virginica</p> <p>Predicted label</p>

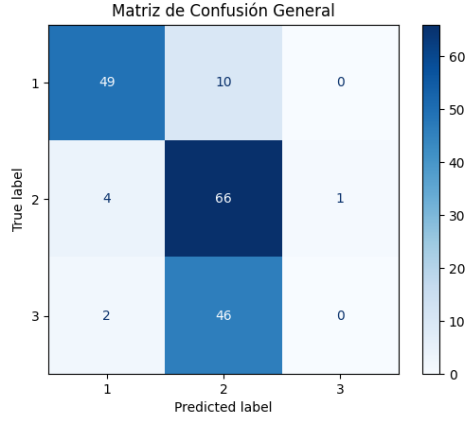
Random Forest	94.66%	<div><p>Matriz de Confusión General</p><table><tr><td>True label \ Predicted label</td><td>setosa</td><td>versicolor</td><td>virginica</td></tr><tr><td>setosa</td><td>50</td><td>0</td><td>0</td></tr><tr><td>versicolor</td><td>0</td><td>47</td><td>3</td></tr><tr><td>virginica</td><td>0</td><td>5</td><td>45</td></tr></table></div>	True label \ Predicted label	setosa	versicolor	virginica	setosa	50	0	0	versicolor	0	47	3	virginica	0	5	45
True label \ Predicted label	setosa	versicolor	virginica															
setosa	50	0	0															
versicolor	0	47	3															
virginica	0	5	45															
SVM (Kernel Lineal)	96%	<div><p>Matriz de Confusión General</p><table><tr><td>True label \ Predicted label</td><td>setosa</td><td>versicolor</td><td>virginica</td></tr><tr><td>setosa</td><td>50</td><td>0</td><td>0</td></tr><tr><td>versicolor</td><td>0</td><td>47</td><td>3</td></tr><tr><td>virginica</td><td>0</td><td>3</td><td>47</td></tr></table></div>	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															
SVM (Kernel Gaussiano)	93.33%	<div><p>Matriz de Confusión General</p><table><tr><td>True label \ Predicted label</td><td>setosa</td><td>versicolor</td><td>virginica</td></tr><tr><td>setosa</td><td>50</td><td>0</td><td>0</td></tr><tr><td>versicolor</td><td>0</td><td>46</td><td>4</td></tr><tr><td>virginica</td><td>0</td><td>6</td><td>44</td></tr></table></div>	True label \ Predicted label	setosa	versicolor	virginica	setosa	50	0	0	versicolor	0	46	4	virginica	0	6	44
True label \ Predicted label	setosa	versicolor	virginica															
setosa	50	0	0															
versicolor	0	46	4															
virginica	0	6	44															

<div>SVM (Kernel Polinomial)</div> <div>Grado: 4</div>	<div>95.33%</div>	<div><div>Matriz de Confusión General</div><table><tr><th>True label \ Predicted label</th><th>setosa</th><th>versicolor</th><th>virginica</th></tr><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>3</td><td>47</td></tr></table></div>	True label \ Predicted label	setosa	versicolor	virginica	setosa	50	0	0	versicolor	0	46	4	virginica	0	3	47
True label \ Predicted label	setosa	versicolor	virginica															
setosa	50	0	0															
versicolor	0	46	4															
virginica	0	3	47															

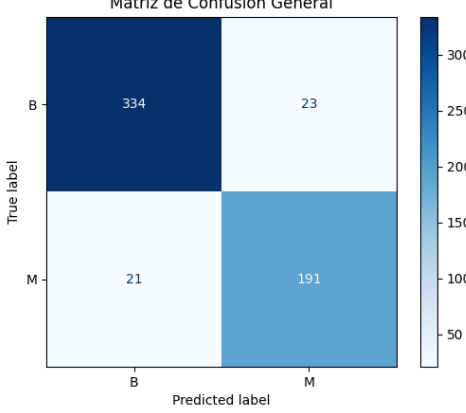
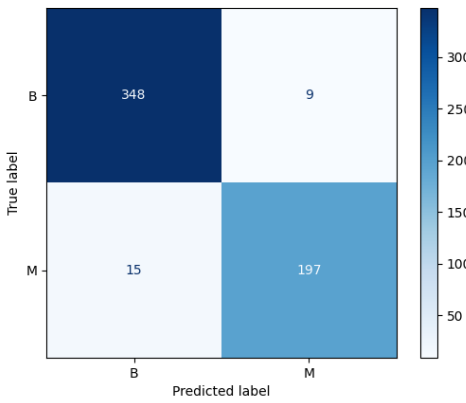
2) Wine

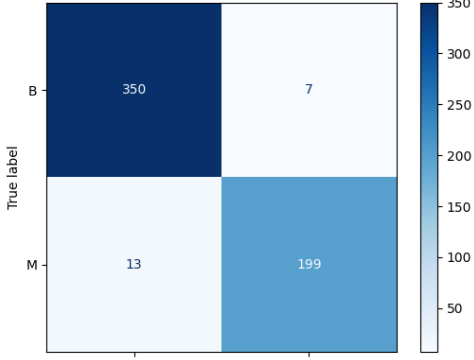
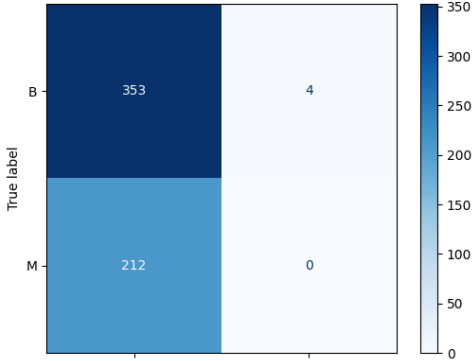
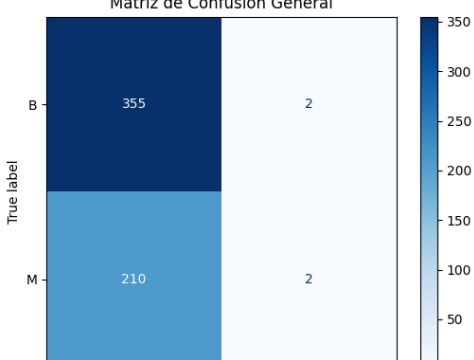
Clasificador	Accuracy	Matriz de Confusión
Árbol de Clasificación J48	85.45%	Matriz de Confusión General
		

Random Forest	96.66%	<p>Matriz de Confusión General</p>  <table><tr><th>True label \ Predicted label</th><th>1</th><th>2</th><th>3</th></tr><tr><th>1</th><td>57</td><td>2</td><td>0</td></tr><tr><th>2</th><td>1</td><td>68</td><td>2</td></tr><tr><th>3</th><td>0</td><td>1</td><td>47</td></tr></table>	True label \ Predicted label	1	2	3	1	57	2	0	2	1	68	2	3	0	1	47
True label \ Predicted label	1	2	3															
1	57	2	0															
2	1	68	2															
3	0	1	47															
SVM (Kernel Lineal)	93.30%	<p>Matriz de Confusión General</p>  <table><tr><th>True label \ Predicted label</th><th>1</th><th>2</th><th>3</th></tr><tr><th>1</th><td>58</td><td>1</td><td>0</td></tr><tr><th>2</th><td>2</td><td>63</td><td>6</td></tr><tr><th>3</th><td>1</td><td>2</td><td>45</td></tr></table>	True label \ Predicted label	1	2	3	1	58	1	0	2	2	63	6	3	1	2	45
True label \ Predicted label	1	2	3															
1	58	1	0															
2	2	63	6															
3	1	2	45															
SVM (Kernel Gaussiano)	54.44%	<p>Matriz de Confusión General</p>  <table><tr><th>True label \ Predicted label</th><th>1</th><th>2</th><th>3</th></tr><tr><th>1</th><td>58</td><td>1</td><td>0</td></tr><tr><th>2</th><td>2</td><td>64</td><td>5</td></tr><tr><th>3</th><td>0</td><td>2</td><td>46</td></tr></table>	True label \ Predicted label	1	2	3	1	58	1	0	2	2	64	5	3	0	2	46
True label \ Predicted label	1	2	3															
1	58	1	0															
2	2	64	5															
3	0	2	46															

<div>SVM (Kernel Polinomial)</div> <div>Grado: 6</div>	<div>63.88%</div>	<div>Matriz de Confusión General</div>  <table><tr><th>True label \ Predicted label</th><th>1</th><th>2</th><th>3</th></tr><tr><th>1</th><td>49</td><td>10</td><td>0</td></tr><tr><th>2</th><td>4</td><td>66</td><td>1</td></tr><tr><th>3</th><td>2</td><td>46</td><td>0</td></tr></table>	True label \ Predicted label	1	2	3	1	49	10	0	2	4	66	1	3	2	46	0
True label \ Predicted label	1	2	3															
1	49	10	0															
2	4	66	1															
3	2	46	0															

3) Breast Cancer

Clasificador	Accuracy	Matriz de Confusión									
<p>Árbol de Clasificación J48</p>	<p>92.27%</p>	<p>Matriz de Confusión General</p>  <table border="1"> <thead> <tr> <th>True label \ Predicted label</th><th>B</th><th>M</th></tr> </thead> <tbody> <tr> <th>B</th><td>334</td><td>23</td></tr> <tr> <th>M</th><td>21</td><td>191</td></tr> </tbody> </table>	True label \ Predicted label	B	M	B	334	23	M	21	191
True label \ Predicted label	B	M									
B	334	23									
M	21	191									
<p>Random Forest</p>	<p>95.78%</p>	<p>Matriz de Confusión General</p>  <table border="1"> <thead> <tr> <th>True label \ Predicted label</th><th>B</th><th>M</th></tr> </thead> <tbody> <tr> <th>B</th><td>348</td><td>9</td></tr> <tr> <th>M</th><td>15</td><td>197</td></tr> </tbody> </table>	True label \ Predicted label	B	M	B	348	9	M	15	197
True label \ Predicted label	B	M									
B	348	9									
M	15	197									

SVM (Kernel Lineal)	96.48%	<p>Matriz de Confusión General</p>  <table border="1"><thead><tr><th></th><th>B</th><th>M</th></tr></thead><tbody><tr><th>B</th><td>350</td><td>7</td></tr><tr><th>M</th><td>13</td><td>199</td></tr></tbody></table>		B	M	B	350	7	M	13	199
	B	M									
B	350	7									
M	13	199									
SVM (Kernel Gaussiano)	62.06%	<p>Matriz de Confusión General</p>  <table border="1"><thead><tr><th></th><th>B</th><th>M</th></tr></thead><tbody><tr><th>B</th><td>353</td><td>4</td></tr><tr><th>M</th><td>212</td><td>0</td></tr></tbody></table>		B	M	B	353	4	M	212	0
	B	M									
B	353	4									
M	212	0									
SVM (Kernel Polinomial) Grado: 5	62.76%	<p>Ver</p> <p>Matriz de Confusión General</p>  <table border="1"><thead><tr><th></th><th>B</th><th>M</th></tr></thead><tbody><tr><th>B</th><td>355</td><td>2</td></tr><tr><th>M</th><td>210</td><td>2</td></tr></tbody></table>		B	M	B	355	2	M	210	2
	B	M									
B	355	2									
M	210	2									

Conclusión

La aplicación de diferentes algoritmos de aprendizaje automático a los conjuntos de datos Iris, Wine y Breast Cancer, utilizando métodos de validación Hold-Out y validación cruzada, ha proporcionado insights valiosos para la selección de modelos en futuros análisis. En el conjunto de datos Iris, todos los modelos, y en particular los SVM con distintos kernels, han mostrado un rendimiento excepcional, destacándose por una precisión perfecta del 100% en la validación Hold-Out, lo que resalta su efectividad para este tipo de datos. Por otro lado, en el conjunto de datos Wine, se observó una variabilidad significativa en el rendimiento; mientras que Random Forest logró una precisión perfecta, los SVM con kernels Gaussiano y polinomial tuvieron una disminución notable en su precisión, sugiriendo la crítica importancia de la selección del kernel en entornos de datos con características específicas. En cuanto al conjunto de datos Breast Cancer, Random Forest mostró un alto rendimiento de manera consistente en ambos métodos de validación, demostrando su robustez y fiabilidad. A su vez, los SVM con kernel lineal mantuvieron un rendimiento sólido en general, aunque los resultados variaron considerablemente con los kernels Gaussiano y polinomial, subrayando la relevancia de una adecuada elección del kernel en función de la naturaleza de los datos. La validación cruzada reveló una ligera disminución en el rendimiento en comparación con Hold-Out para casi todos los modelos, resaltando la importancia de evaluar los modelos bajo diferentes escenarios para comprender mejor su capacidad de generalización.