PRÁCTICA 3:
MICROARRAY DE EXPRESIÓN GENÉTICA

MINERÍA DE DATOS
PATRICIA AGUADO LABRADOR

CONJUNTO DE DATOS AML-ALL-NORMALIZED

Descripción del dataset: conjunto de datos sobre la leucemia formado por 72 instancias con 7129 atributos y una clase con dos posibles valores, AML y ALL.

Tasas de error validación cruzada 10 particiones	
J48	0,2083
NB	0,0138
IBK1	0,1527
REG LOGISTICA	0,125
MLP (H10)	0,0277
SVM (Lineal)	0,0138

J48 es el algoritmo que peor se comporta y puede ser debido al gran número de atributos.

SELECCIÓN DE ATRIBUTOS: FILTRO

En el caso de la elección de 4 y 8 atributos, los métodos llegan a compartir dos a dos, de uno a 2 atributos. En el caso de la elección de 16 vemos que los atributos compartidos no aumentan y por último, en el caso de la elección de 32 vemos que los tres métodos comparten 10 atributos, lo que significa que estos son atributos significativos ya que a la hora de seleccionar los mejores todos los métodos de filtro los han incluido en su selección.
El método CFsubsetEval no termina y esto podría deberse a que existen demasiados atributos en el dataset.

4 atributos:

Incertidumbre Simétrica: 1834, 4847, 1882, 3252, 2288 760, 6041, 6855.
ReliefF: 3252, 4196, 1779, 4847, 2402, 4951, 1834, 1829.
SVM(lineal): 1882, 1834, 1779, 1796, 4196, 5348, 5094, 804.
CFsubsetEval: no terminó.

8 atributos:

Incertidumbre Simétrica: 1834, 4847, 1882, 3252, 2288 760, 6041, 6855.
ReliefF: 3252, 4196, 1779, 4847, 2402, 4951, 1834, 1829.
SVM(lineal): 1882, 1834, 1779, 1796, 4196, 5348, 5094, 804.
CFsubsetEval: no terminó.

16 atributos:

Incertidumbre Simétrica: 1834, 4847, 1882, 3252, 2288, 760, 6041, 6855, 1685, 6376, 2354, 4373, 4377, 4366, 2402, 758.

ReliefF: 3252, 4196, 1779, 4847, 2402, 4951, 1834, 1829, 6041, 2288, 1882, 6201, 1745, 3320, 6919, 2363.
SVM(lineal): 1882, 1834, 1779, 1796, 4196, 5348, 5094, 804, 5107, 3847, 4951, 4847, 2354, 6539, 1933, 2288.
CFsubsetEval: no terminó.

32 atributos:
Incertidumbre Simétrica: 1834, 4847, 1882, 3252, 2288, 760, 6041, 6855, 1685, 6376, 2354, 4373, 4377, 4366, 2402 758, 4328, 1144, 3320, 2642, 2335, 1829, 2128, 6281, 4229, 2020, 1779, 2121, 4196, 1902, 1926, 1400.
ReliefF: 3252, 4196, 1779, 4847, 2402, 4951, 1834, 1829, 6041, 2288, 1882, 6201, 1745, 3320, 6919, 2363, 2111, 4052, 2642, 2121, 1674, 6225, 461, 1249, 4366, 2354, 2020, 6539, 1291, 2546, 1260, 235.
SVM(lineal): 1882, 1834, 1779, 1796, 4196, 5348, 5094, 804, 5107, 3847, 4951, 4847,
2354, 6539, 1933, 2288, 6041, 2300, 3252, 129, 6154, 1941, 2475, 6225, 3320, 3714, 2111, 134, 1975, 6184, 461, 1685.
CFsubsetEval: no terminó.

Tasas de error validación cruzada 10 particiones con selección de atributos				
	Incertidumbre simétrica	**ReliefF**	**Eliminación recursiva con SVM**	**CFsubsetEval**
4 **J48**	0,0972	0,0833	0,0833	x
8 **J48**	0,1527	0,1111	0,0833	x
16 **J48**	0,1527	0,1527	0,125	x
32 **J48**	0,1388	0,1666	0,1527	x

Vemos que el método de árboles, J48, funciona mejor cuanto menor es el número de atributos significativos seleccionados. De los tres métodos el que mejor funciona con J48 es el de la selección de atributos con SVM.

Tasas de error validación cruzada 10 particiones con selección de atributos				
	Incertidumbre simétrica	**ReliefF**	**Eliminación recursiva con SVM**	**CFsubsetEval**
4 **NB**	0,0555	0,0833	0,0277	x
8 **NB**	0,0555	0,0277	0,0277	x
16 **NB**	0,0416	0,0555	0	x
32 **NB**	0,0416	0,0416	0,0138	x

Al utilizar el algoritmo de Naive Bayes para clasificar vemos que las tasas de error mejoran al utilizar los métodos de incertidumbre y SVM con mayor número de atributos.

Tasas de error validación cruzada 10 particiones con selección de atributos				
	Incertidumbre simétrica	**ReliefF**	**Eliminación recursiva con SVM**	**CFsubsetEval**
4 **IBK1**	0,0833	0,1111	0	x
8 **IBK1**	0,0694	0,0555	0,0138	x
16 **IBK1**	0,0416	0,0694	0	x
32 **IBK1**	0,0416	0,0694	0	x

El algoritmo de vecinos próximos funciona muy bien con el algoritmo SVM y con incertidumbre simétrica con bastantes atributos. En el caso de la selección de atributos funciona mejor con un menor número de atributos.

Tasas de error validación cruzada 10 particiones con selección de atributos				
	Incertidumbre simétrica	**ReliefF**	**Eliminación recursiva con SVM**	**CFsubsetEval**
4 **R.LOGIST**	0,0694	0,0555	0	x
8 **R.LOGIST**	0,0555	0,0972	0,0138	x
16 **R.LOGIST**	0,0417	0,0694	0	x
32 **R.LOGIST**	0,0416	0,0416	0	x

Al utilizar el algoritmo de regresión logística vemos que todos los métodos de selección de atributos funcionan muy bien cuanto mayor es el número de atributos seleccionados.

Tasas de error validación cruzada 10 particiones con selección de atributos				
	Incertidumbre simétrica	**ReliefF**	**Eliminación recursiva con SVM**	**CFsubsetEval**
4 **MLP (H10)**	0,0694	0,0555	0	x
8 **MLP (H10)**	0,0416	0,0694	0	x
16 **MLP (H10)**	0,0138	0,0694	0	x
32 **MLP (H10)**	0,0277	0,0277	0	x

El algoritmo de MLP con 10 capas ocultas funciona muy bien con el algoritmo SVM y con los demás funciona mejor con un mayor número de atributos.

Tasas de error validación cruzada 10 particiones con selección de atributos				
	Incertidumbre simétrica	**ReliefF**	**Eliminación recursiva con SVM**	**CFsubsetEval**
4 **SVM (lineal)**	0,0694	0,0555	0	x
8 **SVM (lineal)**	0,0694	0,0555	0	x
16 **SVM (lineal)**	0,0555	0,0277	0	x
32 **SVM (lineal)**	0,0277	0,0277	0	x

Al utilizar el algoritmo SVM para clasificar vemos que las tasas de error mejoran al utilizar los métodos de incertidumbre y ReliefF con mayor número de atributos, y que obtiene las mejores tasas de error empleando eliminación recursiva con SVM.

DISCUSIÓN DE RESULTADOS:

Todos los métodos de filtro con los que hemos trabajado consiguen que los algoritmos de clasificación obtengan tasas de error bajas, pero los números reflejan que el mejor comportamiento de los algoritmos se obtiene utilizando el filtro de máquinas de soporte vectorial.

SELECCIÓN DE ATRIBUTOS: ENVOLTORIO

Atributos seleccionados:

J48: 4847.
NB: 6, 461, 760, 6615.
IBK1: 28, 1834, 3258, 3549.
Reg. Log: 43, 1882, 6049.
MLP(10):1795, 1834, 2288.
SVM(lineal): 162, 1796, 2111, 3252.

Tasas de error validación cruzada 10 particiones						
	J48	**NB**	**IBK1**	**Reg. Log**	**MLP(10)**	**SVM(lineal)**
J48	0,0555					
NB		0,0138				
IBK1			0			
Reg. Log				0,0138		
MLP(10)					0,0555	
SVM(lineal)						0,0277

Vemos que al utilizar los métodos de envoltorio los algoritmos que peor se comportan con este conjunto de datos son J48 y MLP, aunque obtenemos porcentajes de error del 5.5%, una cantidad bastante pequeña.
K vecinos más próximos es el que mejor se comporta, seguido de NB y regresión logística.

COMPARACIÓN CON MÉTODOS DE FILTRO:

En cuanto a los atributos seleccionados vemos que la mayoría ahora seleccionados ya lo habían sido cuando se aplicaron los métodos de filtro como por ejemplo el atributo 1834, el atributo 4847, el atributo 760, etc.

En cuanto a las tasas de error obtenidas:

Al emplear J48 con métodos de envoltorio obtenermos una tasa de error mucho menor que cuando empleábamos este algoritmo para clasificar el conjunto de datos junto con métodos de filtro.

Al emplear el resto de algoritmos obtenemos tasas de error muy pequeñas al utilizar métodos de envoltorio pero estas pueden conseguirse empleando Eliminación recursiva con SVM para la selección de atributos junto con los algoritmos de clasificación pertinentes, e incluso en ocasiones podemos obtener una tasa de error nula.

PCA

El algoritmo de clasificación que menor tasa de error obtiene al aplicarlo sobre el conjunto de datos AML-ALL-normalized es IBK1 (M1 = IBK). El número de atributos significativos que seleccionaba este método era 4, por tanto, n=4.

El tiempo de cómputo a la hora de realizar el análisis de componentes principales con el conjunto de datos principal superaba las 24h, no he conseguido hacer que termine por lo que no puedo realizar los ejercicios de PCA.

CONCLUSIONES

Por lo que hemos podido ver los métodos de envoltorio nos proporcionan tasas de error muy buenas pero el problema es que son propensos al sobreajuste al utilizar el método de aprendizaje como evaluador.

El método de eliminación recursiva con SVM en métodos de filtro nos ha facilitado las mejores tasas de error de media con cualquier algoritmo utilizado para la clasificación con validación cruzada.