PRÁCTICA 2:
SELECCIÓN DE SUBCONJUNTO DE ATRIBUTOS

MINERÍA DE DATOS
PATRICIA AGUADO LABRADOR

EJERCICIO 1

CfsSubsetEval (Forward)	Árboles de decisión	OneR	ReliefF	Recursive SVMfeature Evaluation	Incertidumbre simétrica
LSHM	LSHM	Cumulante1	USHM	USHM	Cumulante1
USHM	USHM	Momento1	LSHM	Cumulante4	Momento1
Momento1	Momento1	Cumulante3	Cumulante3	Momento3	Cumulante6
Momento4	Momento2	Cumulante6	Momento4	LSHM	Cumulante3
Cumulante1	Momento3	USHM	Cumulante4	Momento4	LSHM
Cumulante6	Momento4	LSHM	Cumulante6	Cumulante3	USHM
Skewness	Cumulante6	Kurtosis	Momento2	FactorCresta	Kurtosis
Kurtosis	Kurtosis	FactorForma	Cumulante2	Cumulante6	Momento4
xr	Skewness	Momento2	Axm	Skewness	Cumulante4
	FactorForma	xr	Xp	Xp	Momento2

Podemos ver que con cualquiera de los métodos utilizados hay tres de los atributos que se repiten en cada una de las selecciones (LSHM, USHM, Cumulante6), por lo que podemos pensar que tienen gran relevancia en el conjunto de atributos.
Al utilizar el método de selección de atributos basada en la correlación con las parametrizaciones por defecto, vemos que este sólo selecciona nueve atributos.

	Tasa de error 10-XV con todos los atributos	Tasa de error 10-XV con los atributos seleccionados					
Algoritmo		**Árboles de decisión**	**OneR**	**ReliefF**	**Recursive SVMfeature Evaluation**	**Incertidumbre simétrica**	**CfsSubsetEval**
J48	0,1953	0,1813	0,1906	0,2232	0,2186	0,1674	0,1674
NB	0,4744	0,5162	0,3813	0,3813	0,5255	0,3627	0,4232
IBK3	0,3813	0,4046	0,3581	0,3627	0,3813	0,3581	0,3674
R.Logística	0,2465	0,372	0,4139	0,2325	0,2093	0,2372	0,372
MLP (H10)	0,3813	0,386	0,3813	0,3674	0,2744	0,3116	0,3441
SVM (lineal)	0,386	0,3813	0,4	0,4	0,3813	0,3906	0,3953

Analizando las tasas de error de la tabla podemos observar que al utilizar el método de cross-validation, por lo general, el algoritmo J48, basado en la construcción de árboles, se comporta bien tanto con todo el conjunto de atributos como con cualquier algoritmo de selección de atributos.

Vemos que obtenemos tasas de error similares al no realizar la selección de atributos y al realizarla mediante el método basado en la correlación (CfsSubsetEval), por lo que podríamos decir que utilizar este método con la idea de eliminar atributos irrelevantes no tiene mucho efecto en el porcentaje de aciertos de la clasificación con los diferentes algoritmos de aprendizaje.

EJERCICIO 2

CfsSubsetEval Greedy Stepwise (forward) Por defecto	CfsSubsetEval Greedy Stepwise (forward) 10 atributos
LSHM	Momento1
USHM	Cumulante6
Momento1	LSHM
Momento4	USHM
Cumulante1	Cumulante1
Cumulante6	Kurtosis
Skewness	Momento4
Kurtosis	Skewness
Xr	Cumulante3
	FactorForma

El subconjunto obtenido al forzar la selección de diez atributos, no contiene los atributos del subconjunto obtenido con las configuraciones por defecto ya que no se encuentra el atributo 'Xr' entre ellos. Esto podría deberse a un problema de precisión en el algoritmo de selección ya que el atributo que se encontraba entre los nueve mejores ahora no se encuentra entre los diez.

EJERCICIO 3

WrapperSubsetEval + GreedyStepwise J48	WrapperSubsetEval + GreedyStepwise SVM (linea)
Momento1	USHM
Cumulante6	LSHM
Momento4	Momento4
USHM	Cumulante4
Kurtosis	Momento1
Cumulante4	Cumulante1
xr	Cumulante6
Cumulante3	Momento3
Xp	Skewness
Cumulante1	Kurtosis

Al realizar la selección de diez atributos con estos dos métodos nos encontramos de nuevo con los atributos USHM y Cumulante6, los cuales ya obteníamos en los rankings de los otros métodos utilizados al principio de la práctica. Este hecho otorga más peso a la suposición de que ambos son atributos relevantes.

	Tasa de error 10-XV con todos los atributos	Tasa de error 10-XV con los atributos seleccionados						
Algoritmo		**Árboles de decisión**	**OneR**	**ReliefF**	**Recursive SVMfeature Evaluation**	**Incertidumbre simétrica**	**CfsSubsetEval**	**WrapperSubsetEval + GreedyStepwise**
J48	0,1953	0,1813	0,1906	0,2232	0,2186	0,1674	0,1674	0,1348
NB	0,4744	0,5162	0,3813	0,3813	0,5255	0,3627	0,4232	
IBK3	0,3813	0,4046	0,3581	0,3627	0,3813	0,3581	0,3674	
R.Logística	0,2465	0,372	0,4139	0,2325	0,2093	0,2372	0,372	
MLP (H10)	0,3813	0,386	0,3813	0,3674	0,2744	0,3116	0,3441	
SVM (lineal)	0,386	0,3813	0,4	0,4	0,3813	0,3906	0,3953	0,3813

Vemos que el método de envoltorio utilizado junto con el algoritmo J48 mejora mucho las tasas de error obtenidas hasta el momento, conseguimos tener solo un 13.48% de error en la clasificación. Por otro lado, el método de envoltorio utilizado con SVM de kernel lineal obtiene una tasa parecida al resto de métodos de selección de atributos.