Trabajo 1 - Aprendizaje supervisado

Para este trabajo se ha seleccionado un dataset de intención de voto en Estados Unidos de 1984 (vote.arff), este conjunto de datos incluye 16 atributos que describen el voto afirmativo, en contra o abstención de distintos temas políticos relevantes de la época. El atributo objetivo es el partido político de cada votante, republicano o demócrata.

Es importante tener en cuenta que hay 267 (54.8%) instancias de votantes demócratas y 168 (45.2%) instancias de votantes republicanos, por lo que los resultados pueden estar sesgados al no trabajar con clases balanceadas.

En primer lugar, se ha aplicado el algoritmo de **máquinas de soporte de vectores** (SVM, functions.LibSVM en Weka), estableciendo una validación cruzada con 20 subconjuntos. Este clasificador se encarga de hallar un hiperplano en la dimensión de los atributos, en este caso de dimensión 16, para separar los ejemplos de las distintas clases.

La matriz de confusión obtenida con este modelo es la siguiente:

	Clasificado Demócrata	Clasificado Republicano
Es Demócrata	253	14
Es Republicano	5	163

Se puede observar que se han clasificado correctamente 416 instancias, fallando en las 19 restantes, esto supone un error medio del 4.37%.

La media ponderada de la precisión (las etiquetas predichas correctamente entre el conjunto de etiquetas predichas) de ambas clases es del 95.8%, mientras que la media ponderada del recall (las etiquetas predichas correctamente dentro del conjunto total de las instancias) de ambas clases es del 95.6%. Esto resulta en un valor de F1 del 95.7%.

El valor de Kappa, que es la probabilidad de que una instancia se clasifique de la misma forma en dos ejecuciones del modelo, es igual al 90.88%.

El área bajo la curva ROC, que evalúa la calidad del algoritmo clasificador, siendo el área de un clasificador perfecto 1 y 0.5 el de uno aleatorio, es igual a 0.94.

Se puede determinar que aplicando SVM sobre este conjunto de datos se obtiene un modelo clasificador adecuado.

A continuación se aplicará el algoritmo **random forest** (trees.RandomForest en Weka), estableciendo otra vez una validación cruzada con 20 subconjuntos. En este caso, el modelo construye un conjunto de hipótesis sobre los atributos para combinar las predicciones de alguna forma con el objetivo de mejorar la clasificación de los ejemplos.

La matriz de confusión obtenida con este modelo es la siguiente:

	Clasificado Demócrata	Clasificado Republicano
Es Demócrata	261	6
Es Republicano	9	159

En este caso se han clasificado correctamente 420 instancias y ha fallado en las 15 restantes, esto es un error medio del 3.45%. Este modelo ha clasificado erróneamente menos instancias que el anterior, aunque ha clasificado erróneamente más ejemplos demócratas que en realidad eran republicanos.

Las métricas de este modelo se describen a continuación, entre paréntesis se muestran los valores del modelo anterior, en todo caso se da el valor de la media ponderada de la métrica entre ambas clases:

Precisión: 96.5% *(95.8%)*

Recall: 96.6% *(95.6%)*

F1: 96.5% *(95.7%)*

Kappa: 92.7% *(90.88%)*

Área ROC: 0.993 *(0.94)*

Como se puede observar, el modelo basado en random forest es superior al basado en máquinas de soporte de vectores en todas las métricas analizadas, alcanzando un rendimiento casi perfecto.