Trabajo Práctico 4: Métodos supervizados avanzados

1. Evalúe el efecto de la complejidad del clasificador en boosting. Use los dos datasets adjuntos que se detallan. Un dataset es el problema de las espirales anidadas con ruido, el otro es el conocido problema "diagonal". Controle la complejidad de los arboles poniendo un máximo a la profundidad de los mismos (ejemplo debajo). Utilice 200 árboles para cada ensemble. Estime el error de clasificación en test en función de la profundidad máxima para valores de ésta de 1 a 20.
Haga una gráfica y analice el resultado en cada caso.

```
#Ejemplo de boosting con xgboost sobre espirales:
require(xgboost)
XGB.nrounds=500   #total de arboles/ciclos
XGB.eta=0.1           #learning rate, similar a redes
XGB.max.depth=3   #profundidad maxima de los arboles, complejidad

x.train<-as.matrix(esp_train[,1:2])
y.train<-esp_train[,3]

m.xgb <- xgboost(data=x.train, label=as.integer(y.train)-1, objective="binary:hinge", nrounds=XGB.nrounds,
early_stopping_rounds=1000, eta=XGB.eta, max.depth=XGB.max.depth, colsample_bytree=1, verbose=0,
subsample=1)
```

2. Evalúe el efecto de la cantidad de features evaluadas a cada paso para Random Forest (parametro mtry en R). Use el dataset RRL que está adjunto al texto (variable a predecir es "Tipo"), cambiando mtry como fracción del total de features en potencias de 1/2 (en este caso, 69, 34, 17, 8, 4, 2, 1). Utilice 1000 árboles para cada ensemble. Estime el error de clasificación OOB en función de la fracción utilizada, como promedio de 5 corridas para cada valor de mtry. Haga una gráfica y analice el resultado.

3. Aplicación a datos anchos. Compare los resultados de clasificación de RandomForest (librería randomForest), xgboost con árboles (librería xgboost) y SVM con kernels RBF y Polinomial (librería e1071). Use el dataset lampone, para predecir la clase (variable n_tipo), con una metodología adecuada para seleccionar los parámetros internos y estimar el error.

4. Opcional (2 puntos) Repita la comparación del ejercicio anterior para el dataset RRL que está incluido en el archivo del punto 1. La variable a predecir es "Tipo". Los datos son observaciones de estrellas, el objetivo es determinar si es una estrella variable o no.

Entregue un notebook con todo el código y los comentarios correspondientes.