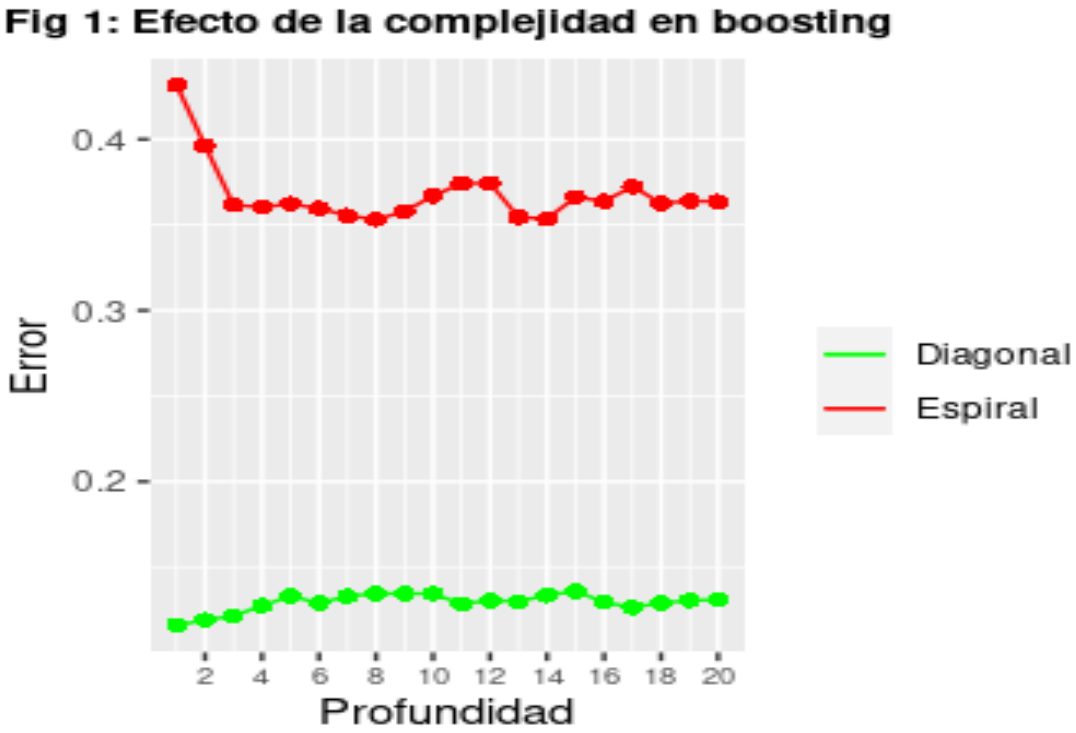
Práctica 4: Métodos supervizados avanzados

Alumno: Pablo Alonso

Ej 1)

A continuación una gráfica con los resultados obtenidos:



En el problema **Espiral** se observa un menor error de ajuste con árboles de una profundidad 8, mientras que en el **Diagonal**, el menor error se encuentra con árboles de profundidad 1. Es decir que **Espiral** requiere de un boosting con árboles más complejos que **Diagonal**para minimizar el error. Esto tiene sentido al ser **Espiral** un problema más complejo al estar las espirales anidadas, mientras que **Diagonal** se trata de 2 distribuciones cuyos centros están separados.

La gráfica da una idea entre el trade-off entre sesgo y varianza, pues para el caso **Espiral** el error desciende hasta llegar a una profundidad 8 pero tiende a subir para árboles más profundos, osea para los que tienen un mayor error de

varianza. En conclusión, usando la misma cantidad de iteraciones de boosting para el problema **Espiral** y **Diagonal**, **Diagonal** requiere de árboles con más sesgo que **Espiral**.

Ej 2)

Lo primero es que hay muchas variables y pocos datos. Es decir que los datos como puntos en su espacio de features se alejan entre si provocando que haya varianza entre las soluciones. La siguiente tabla muestra los resultados obtenidos:

Algoritmo	Error con 5 folds
SVM kernel Polynomial	0.1
SVM kernel Gaussiano	0.25
Ada Boosting	0.078
Random forest	0.078

Los resultados anteriores son demasiados optimistas para los algoritmos ensembles y para SVM con kernel polynomial teniendo en cuenta que los algoritmos usados tienen parámetros libres y pueden ser muy flexibles, además del problema de la dimensionalidad en los datos. Al hacer un 5 fold se observa demasiada varianza en el error de test. Por ejemplo para Adaboost los errores de test para 5 fold fueron 0.18, 0.11, 0.1, 0, 0 y la media es 0.078.

Ej 3)

Algoritmo	Error con 5 folds
SVM kernel Polynomial	0.09
SVM kernel Gaussiano	0.08
Ada Boosting	0.038
Random forest	0.04

Los algoritmos ensembles presentan una mejora notable con respecto a los no ensembles. Es decir, para este problema se obtiene una mejor performance mediante un conjunto de clasificadores con sesgo (Adaboost) o sin sesgo (Random Forest) que usando clasificadores individuales sin sesgo (las smvs). En parte esto es debido a que los árboles de los métodos ensembles se especializan en diferentes subconjuntos de los datos, haciendo que la respuesta devuelta por el clasificador final sea más robusta puesto que es una ponderación de las respuestas de varios clasificadores y esto también la hace más estable. Además, las svms tienden a sobreajustar mientras que los ensembles buscan el punto óptimo entre sesgo y varianza.

En el experimento se comprueba que el error de Adaboost puede ser menor que el de Random Forest debido a que Adaboost implementa mejoras teóricas con respecto a Random Forest (boostraps con distintos pesos y reducción de una función de costo en cada paso) . Sin embargo los resultados de Random Forest son muy buenos sin haber necesitado ajustar ningún parámetro (para Adaboost el parámetro a ajustar fue la complejidad de los árboles).