



Práctica 2 - Clasificación

La práctica 2 de la asignatura *Machine Learning* consiste en la resolución de un problema de **Clasificación**. El problema a resolver se describe a continuación.

Requisitos de la práctica

La práctica consta de los siguientes entregables:

1. Una libreta de Python en la cual se realizará el desarrollo de la práctica. La libreta incluirá todo el código de las operaciones realizadas para el preprocessamiento, entrenamiento y validación de los modelos, así como **una sección de conclusiones** en la cual se interpretarán no solo las medidas de calidad obtenidas por el modelo sino también el modelo en sí mismo. Para facilitar el proceso de corrección, **todos los datos deberán ser cargados desde una URL externa**, y no desde el almacenamiento local de la libreta utilizada.
2. Un vídeo corto de alrededor de 15 minutos en el cual se presenten los desarrollos, resultados y conclusiones obtenidas para el problema resuelto. Para la entrega del vídeo, se subirá a la nube y se entregará el enlace.

Será imprescindible realizar los siguientes procesos durante la resolución del problema:

- Preprocesamiento adecuado del conjunto de datos.
- Entrenamiento y validación de diferentes *ensembles* para clasificación. No se deberán evaluar algoritmos de clasificación independientes. **Todas las soluciones propuestas deberán ser con ensembles.**
- Optimización con Grid Search de los hiperparámetros existentes.
- Validación y comparación de los modelos mediante la medida de calidad *accuracy* empleando el **conjunto de test proporcionado**. Se podrán emplear otras medidas de calidad (*precision*, *recall*, matriz de confusión, *AUC*, etc.) con el fin de extraer las conclusiones oportunas sobre el funcionamiento del modelo propuesto.
- De cara a evaluar justamente, la semilla para todos aquellos métodos estocásticos será: `random_state=1337`.

Descripción del problema

Disponemos de un conjunto de datos con imágenes de escenas de la naturaleza de todo el mundo.

El problema de clasificación a resolver consiste en determinar la etiqueta de una imagen conteniendo una escena entre un conjunto de etiquetas predefinidas.

El conjunto de datos

El conjunto de datos contiene 17.000 imágenes de tamaño 150×150 píxeles etiquetadas en 6 categorías diferentes:

- `edificios -> 0`
- `bosques -> 1`
- `glaciares -> 2`
- `montañas -> 3`
- `mares -> 4`
- `calles -> 5`

Los conjuntos de entrenamiento y test se han separado en dos directorios diferentes comprimidos con extensión `.zip`. El conjunto de entrenamiento dispone de 14.000 imágenes y el directorio de test dispone de 3.000

imágenes. Las imágenes han sido agrupadas en subdirectorios en función de su etiqueta.