Ciencias de Datos 2025 - Parcialito 3

El objetivo de este trabajo es aplicar métodos de clasificación supervisada para predecir el tipo de cobertura forestal en base a variables cartográficas, utilizando el conjunto de datos **Covertype**. Se evaluarán distintos modelos, su desempeño comparativo y la interpretación de métricas de clasificación.

Descripción del conjunto de datos

Se utilizará el conjunto de datos **Covertype**, cuyo objetivo es predecir el tipo de cobertura forestal utilizando únicamente variables cartográficas (sin datos obtenidos por sensores remotos). La cobertura forestal real para cada observación (una celda de 30 x 30 metros) fue determinada por el sistema RIS (Resource Information System) del Servicio Forestal de los EE. UU. (USFS), Región 2. Las variables independientes provienen de datos originales del Servicio Geológico de los EE. UU. (USGS) y del USFS. Los datos se encuentran en formato crudo (no escalados) y contienen columnas binarias (0 o 1) para variables cualitativas como las áreas silvestres y los tipos de suelo.

El área de estudio incluye cuatro zonas silvestres ubicadas en el Bosque Nacional Roosevelt, en el norte del estado de Colorado. Estas regiones representan áreas forestales con mínimas perturbaciones humanas, por lo que los tipos de cobertura forestal presentes son principalmente el resultado de procesos ecológicos naturales, independientes de prácticas de manejo forestal.

Información adicional sobre las zonas silvestres:

- **Neota (área 2)** probablemente tiene la altitud media más alta de las cuatro.
- **Rawah (área 1)** y **Comanche Peak (área 3)** tienen altitudes medias intermedias.
- **Cache la Poudre (área 4)** presenta la altitud media más baja.

Especies arbóreas predominantes:

- **Neota**: abeto/abetos alpinos (tipo 1).
- **Rawah y Comanche Peak**: pino de lodgepole (tipo 2), seguido de abeto/abetos alpinos y álamo temblón (tipo 5).
- **Cache la Poudre**: pino ponderosa (tipo 3), abeto de Douglas (tipo 6) y álamo/cauce de río (tipo 4).

Rawah y Comanche Peak son más representativos del conjunto de datos en general, por su diversidad de especies y valores de las variables predictoras. Cache la Poudre resulta ser más particular por su baja altitud y su composición específica de especies.

La tarea consiste en construir un clasificador para predecir la clase mayoritaria de la variable **Cover_Type** en función de las demás variables.

Ejercicios

► **1.**

(a) **Carga y preprocesamiento del conjunto de datos:**

- Explique la adaptación de los datos para convertirlo en un problema de clasifiacción binaria.
- Inspeccione la distribución de clases.
- Divida el conjunto de datos en entrenamiento y prueba.
- Aplique normalización o estandarización si corresponde.
- Aplique reducción de dimensionalidad, si corresponde.

(b) **Entrenamiento de modelos de clasificación:**

- Entrene al menos **tres modelos** distintos, seleccionados entre los siguientes:
  - Regresión logística
  - `LinearSVC`
  - `SVC` con kernel RBF
  - `NuSVC`
  - `SGDClassifier`
- Para al menos **uno de los modelos**, explore distintos valores de hiperparámetros (por ejemplo: `C`, `gamma`, `nu`, `alpha`).
- Genere visualizaciones para interpretar los valores óptimos de los hiperparámetros.

(c) **Evaluación del desempeño:**

- De la lista que sigue, discuta cuales son las mejores métricas para evaluar los modelos. Seleccione tres.
  - Matriz de confusión

- Curvas ROC y área bajo la curva (AUC)
- Curvas precision–recall
- F1-score
- Índice de Youden
- Kappa de Cohen

- Evalúe cada modelo utilizando las métricas elegidas
- Interprete los resultados obtenidos. Analice los compromisos entre diferentes variables, en casos en donde el cambio de umbral imlpica que una métrica mejora y ota empeora. Discuta la utilidad de cada métrica.

(d) **Comparación y discusión:**

- Compare el desempeño de los modelos.
- Justifique cuál sería el modelo más adecuado para este problema, fundamentado en los resultados y en el contexto del dataset.

(e) **Código y reproducibilidad:**

- Entregue un solo archivo que sea un notebook en donde el código debe ser ejecutable y debe estar documentado.
- Incluya comentarios que expliquen las decisiones tomadas durante el desarrollo.

Entrega: 22 de mayo, por el aula virtual