Tarea 1: — Métodos Estadísticos para el Manejo de Grandes Volúmenes de Datos

Profesor: Pedro Luiz Ramos — Pontificia Universidad Católica de Chile
Ayudante: Diego Andrés Bernal Soto — Facultad de Matemática

Ejercicio: Trabajaremos con el conjunto **Wine Data** del repositorio UCI (`https://archive.ics.uci.edu/ml/datasets/wine`). El archivo contiene 178 observaciones provenientes de tres cultivares de vino tinto de la región de Piamonte (Italia). Para cada muestra se midieron 13 atributos fisicoquímicos y una etiqueta de clase (1, 2 o 3).

El objetivo general es **explorar, reducir la dimensión de los datos y construir modelos de regresión** para predecir el contenido de `Proline` a partir de las demás variables.

Cárgalo directamente, por ejemplo:

```
wine_data <- read.table(
  "https://archive.ics.uci.edu/ml/machine-learning-databases/wine/wine.data",
  sep = ","
)
```

1. **Estadística descriptiva** (1.5 pts)
   - Tipifica todas las variables cuantitativas.
   - Presenta tabla de medias, desviaciones estándar, mínimos y máximos.
   - Incluye un *corrplot*, histogramas y boxplots.
   - Detecta valores atípicos mediante la regla IQR y discútelos brevemente.
2. **Análisis de Componentes Principales (PCA)** (2.0 pts)
   - Aplica PCA sobre las 13 variables estandarizadas (excluye la clase).
   - Retén los componentes necesarios para explicar $\geq 70\,\%$ de la varianza.
   - Interpreta las cargas de los dos primeros componentes y presenta un biplot.
3. **Modelos de regresión para `Proline`** (2.5 pts)
   - Define `Proline` como respuesta y las 12 variables restantes como predictoras.
   - Divide el conjunto en 80 % entrenamiento y 20 % validación (usa `set.seed`).
   - Ajusta:
     a) Regresión lineal ordinaria (OLS).
     b) Lasso (penalización L1).
     c) Ridge (penalización L2).
   - Selecciona el hiperparámetro de penalización en Lasso y Ridge usando el conjunto de *validación*.
   - Reporta el error cuadrático medio de validación (MSE) de cada modelo y discute cuál tuvo mejor desempeño y por qué.

- Menciona las variables más relevantes según los coeficientes finales del Lasso.

Nota sobre la variable respuesta. Se selecciona `Proline` (contenido de prolina en mg/L) como variable a predecir porque: (i) es continua y muestra variabilidad suficiente para aplicar modelos de regresión; (ii) no es linealmente redundante con el resto de los predictores, lo que permite capturar relaciones informativas; y (iii) desde el punto de vista enológico, la prolina está asociada a la madurez y calidad del vino, de modo que estimarla resulta relevante y motivador.

Entrega. Envía: (i) código reproducible en R o Python, (ii) informe en PDF con tablas, gráficos y discusión de los hallazgos.

Descripción de las variables

Variable	Descripción
`Class`	Cultivar de origen (1, 2, 3)
`Alcohol`	Contenido de alcohol (%)
`Malic`	Ácido málico (g/L)
`Ash`	Cenizas (g/L)
`Alcalinity`	Alcalinidad de las cenizas (meq NaOH)
`Magnesium`	Magnesio (mg/L)
`Total_phenols`	Fenoles totales (g/L)
`Flavanoids`	Flavonoides (g/L)
`Nonflav_phenols`	Fenoles no flavonoides (g/L)
`Proanthocyanins`	Proantocianidinas (g/L)
`Color_intensity`	Intensidad de color (u.a.)
`Hue`	Matiz (ratio 420/520 nm)
`OD280/OD315`	OD_{280}/OD_{315} vinos diluidos
`Proline`	Prolina (mg/L) — **respuesta**