Trabajo Práctico 7A: Introducción a ML

Statistical Learning es otro nombre con el que se hace referencia a *Machine Learning* (ML). Lea el segundo capítulo del siguiente libro y responda las preguntas. Lectura recomendada hasta la página 42. A partir de allí hay una sección que introduce el lenguaje `R`, que es opcional:

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An introduction to statistical learning: with applications in R* (2nd ed., corrected printing 2023). Springer.

1. En cada uno de los siguientes ejercicios, indique si en general se espera que un método de aprendizaje de máquinas flexible se comporte mejor o peor que uno inflexible. Justifique su respuesta.

   a) El tamaño de la muestra n es extremadamente grande, y el número de predictores p es pequeño.

   b) El número de predictores p es extremadamente grande, y el número de observaciones n es pequeño.

   c) La relación entre los predictores y la variable dependiente es altamente no lineal.

   d) La varianza de los términos de error, $\sigma^2 = \text{Var}(\epsilon)$, es extremadamente alta.

2. Explique si cada escenario representa un problema de clasificación o de regresión, e indique si el interes principal es inferir o predecir. Especifique n (cantidad de observaciones) y p (cantidad de predictores) en cada caso.

   a) Se recopila un conjunto de datos sobre las 500 empresas más importantes de Estados Unidos. Para cada una de las empresas se registran las ganancias, el número de empleados, la industria y el salario del director ejecutivo. Se tiene interés en comprender qué factores afectan el salario de los directores ejecutivos.

   b) Se está considerando lanzar un nuevo producto y se desea saber si será un éxito o un fracaso. Se recolectan datos de 20 productos similares que fueron lanzados previamente. Para cada producto se ha registrado si fue un éxito o un fracaso, el precio cobrado por el producto, el presupuesto de marketing, el precio de la competencia, y otras diez variables.

   c) Se tiene interes en predecir el % de cambio en el tipo de cambio USD/Euro en relación a los cambios semanales en los mercados de valores mundiales. Para eso se recolectan datos semanalmente durante todo el 2021. Para cada semana se registran el % de cambio de USD/Euro, el % de cambio en el mercado estadounidense, el % de cambio en el mercado británico, y el % de cambio en el mercado alemán.

3. ¿Cuáles son las ventajas y desventajas de un enfoque muy flexible (versus uno menos flexible) para la regresión o clasificación? ¿Bajo qué circunstancias podría preferirse un enfoque más flexible a uno menos flexible? ¿Cuándo podría preferirse un enfoque menos flexible?

4. Describa las diferencias entre un enfoque paramétrico y uno no paramétrico. ¿Cuáles son las ventajas y desventajas de un enfoque paramétrico para regresión o clasificación, a diferencia de un enfoque no paramétrico?

5. La siguiente tabla muestra un conjunto de entrenamiento que consta de seis observaciones, tres predictores, y una variable dependiente cualitativa.

Obs.	X_1	X_2	X_3	Y
1	0	3	0	Rojo
2	2	0	0	Rojo
3	0	1	3	Rojo
4	0	1	2	Verde
5	-1	0	1	Verde
6	1	1	1	Rojo

Suponga que se quiere usar este dataset para predecir Y cuando $X_1 = X_2 = X_3 = 0$ usando K vecinos más cercanos.

a) Calcule la distancia Euclidiana entre cada observación y el punto de prueba $X_1 = X_2 = X_3 = 0$.

b) ¿Cuál es la predicción con $K = 1$? Justifique.

c) ¿Cuál es la predicción con $K = 3$? Justifique.

d) Si el límite de decisión de Bayes en este problema es altamente no lineal, ¿se espera que el mejor valor para K sea grande o pequeño? ¿Por qué?

6. Forma de entrega:

a) Dentro del repositorio `ia-uncuyo-2025`, crear una carpeta con el nombre `tp7-ml`.

b) Dentro de la carpeta `tp7-ml` colocar un archivo con el nombre `tp7A-intro-reporte.md` con las respuestas de los ejercicios.