Guía de Explicación del Código de Métodos Supervisados en R

Este documento explica paso a paso cada sección del código utilizado para crear, entrenar y evaluar modelos supervisados (Regresión Logística y RandomForest) usando descriptores moleculares y actividad biológica simulada.

1. Generación del Dataset

En esta parte se simulan valores de descriptores moleculares típicos (MW, LogP, TPSA, HBD, HBA). Se define una "regla oculta" mediante una combinación lineal que modela la probabilidad de actividad. La función logística transforma esta combinación en una probabilidad entre 0 y 1. Finalmente, se genera una etiqueta binaria (Activa: Sí/No) usando una distribución binomial.

2. División Entrenamiento/Prueba

El conjunto de datos completo se separa en dos subconjuntos: - 'train' (70%): se usa para entrenar los modelos. - 'test' (30%): se usa para evaluar el desempeño. Esta división permite estimar qué tan bien generaliza el modelo a datos nuevos.

3. Modelo de Regresión Logística Este modelo predice la probabilidad de que una sustancia sea activa en función de los descriptores. La fórmula Activa ~ MW + LogP + TPSA + HBD + HBA indica que todos los descriptores son predictores. La función glm con family=binomial ajusta un modelo logístico. Posteriormente, se generan probabilidades de actividad para el conjunto de prueba y se convierten en predicciones binarias usando un umbral de 0.5. La matriz de confusión compara estas predicciones con las etiquetas reales y permite calcular la exactitud del modelo.

4. Modelo Random Forest Random Forest es un método no lineal basado en múltiples árboles de decisión. Cada árbol vota por una clase y el resultado final es el voto mayoritario. Los parámetros principales son: - ntree: número de árboles construidos. - mtry: número de predictores seleccionados aleatoriamente en cada división del árbol. Luego se generan predicciones para el conjunto de prueba y se calcula la matriz de confusión y la exactitud. La función importance muestra qué variables aportan más al modelo, y varImpPlot genera una gráfica de importancia de variables.

Interpretación General: El flujo completo refleja un proceso típico en aprendizaje automático: 1. Crear o cargar datos. 2. Dividir en entrenamiento y prueba. 3. Ajustar modelos supervisados. 4. Evaluar mediante métricas como exactitud. 5. Interpretar resultados en función del contexto biofarmacéutico.