SUPUESTO TEÓRICO-PRÁCTICO

D3: MODELADO Y ANÁLISIS DE DATOS.

CASO 1: CLASIFICACIÓN DE IMÁGENES DE FAUNA SILVESTRE CON REDES NEURONALES

Introducción:

Se le ha encomendado la tarea de desarrollar un sistema de clasificación de imágenes de fauna silvestre utilizando redes neuronales. Se le proporciona un conjunto de datos que contiene imágenes etiquetadas de diversas especies de animales, tomadas en entornos naturales. Se pretende utilizar redes neuronales para identificar y clasificar con precisión las especies de animales a partir de imágenes.

Preguntas para Evaluación:

Bloque 1: Preprocesamiento de Datos

- Describa brevemente el proceso de preprocesamiento de datos que llevaría a cabo para preparar el conjunto de datos de imágenes de fauna silvestre antes de utilizarlo para entrenar la red neuronal. Detalle cómo dividiría su conjunto de datos.
- ¿Qué técnicas de aumento de datos específicas consideraría para mejorar la calidad del conjunto de datos?
- ¿Cómo abordaría el desequilibrio en la cantidad de imágenes por especie en el conjunto de datos?

Bloque 2: Entrenamiento de la Red Neuronal

- Seleccione una arquitectura de red neuronal adecuada para la clasificación de imágenes de fauna silvestre. Explique su elección y justifíquela.
- ¿Qué métricas de evaluación utilizaría para medir el rendimiento del modelo durante el entrenamiento? ¿Qué estrategias implementaría para evitar el sobreajuste en su modelo?

Bloque 3: Presentación de Resultados

- ¿Qué tipo de métricas daría para la presentación del resultado final de la eficiencia del modelo?
- ¿Qué haría para garantizar la reproducibilidad de los resultados?

SUPUESTO TEÓRICO-PRÁCTICO

D3: MODELADO Y ANÁLISIS DE DATOS.

CASO 2: VARIABILIDAD METABOLÓMICA EN TUMORES DE MAMA EN RAZÓN DE CLASIFICACIONES HISTOLÓGICA, INMUNOHISTOQUÍMICA Y GENÓMICA

Introducción:

Se plantea un estudio de metabolómica de tumores de mama realizado sobre muestras de 150 pacientes con algún tipo de tumor de mama. Los tumores se clasifican de acuerdo a diferentes criterios:

1. **GRADO HISTOLÓGICO**. Se valora una serie de parámetros histológicos (diferenciación tubular, pleomorfismo nuclear o número de mitosis) y se establecen **tres grados (grados I a III)**.

2. **CLASIFICACIÓN INMUNOHISTOQUÍMICA**. Clasificación molecular basada en diversos biomarcadores cuantificados por inmunohistoquímica. Se establecen **cinco categorías (I1--I5)**.

3. **CLASIFICACIÓN GENÓMICA.** Similar a la anterior pero basada en expresión génica, en lugar de proteínas, como es el caso de la inmunohistoquímica. En este caso sólo se pudieron analizar 60 muestras por falta de muestra o por calidad del material genético al extraerlo. **Se establecen cinco categorías (G1--G5).**

En estos tumores es muy difícil tener muestras puras, esto es de 100% tejido tumoral. Por ello, en todas las muestras se ha cuantificado la cantidad de tejido tumoral y el número de células tumorales.

Los grupos no están balanceados en términos de N, pero están suficientemente representados.

La tabla de trabajo:

- **SAMPLE:** código de la muestra
- **QUALITY:** calidad del espectro, siendo 1 menor y 3 mayor
- **AGE:** edad de la paciente
- **SIZE (cm):** tamaño del tumor
- **GRADE:** clasificación por Grado Histológico
- **ER:** receptor de estrógenos
- **PR:** receptor de progesterona
- **Ki67:** marcador de proliferación
- **HER2/NEU:** protooncogen

- **IMMUNO:** clasificación Inmunohistoquímica.
- **GENOMICS:** clasificación Genómica
- **% area T:** porcentaje de área tumoral en la muestra
- **% cel T en area T:** porcentaje de células tumorales en el área tumoral
- **Weight (RMN/mg):** peso de la muestra analizada en mg
- Variables explicativas independientes: 34 **METABOLITOS** (Ala, Ac,…) medidos de manera absoluta en µmoles/g tejido, incluye algunos ratios relevantes

El objetivo principal es identificar Metabolitos con diferencias significativas entre los tumores según las clasificaciones propuestas o bien otras que se sugieran.

Otro objetivo es determinar si existen efectos covariables de otras variables medidas como AGE, SIZE o % área T en diferencias probadas de uno o varios Metabolitos.

En este ejercicio, para una aproximación a los objetivos, se pide describir un plan de análisis para los siguientes bloques.

Preguntas para Evaluación:

BLOQUE 1: Tratamiento inicial de los datos

Describir correctamente la muestra. Análisis univariantes y bivariantes necesarios

- Tamaños muestrales
- Estadísticos básicos
- Distribuciones de probabilidad
- Test clásicos y estudios de potencia

BLOQUE 2: Tratamiento multivariante de los datos: reducción de la información y clasificación según efectos principales

Discutir qué resultados pueden obtenerse según los diferentes enfoques:

- Reducción de la dimensión en el espacio de variables independientes (metabolitos). Caracterización de componentes. Discutir las opciones

- Procesos de clasificación lineal y no lineal. Discriminación entre grupos:
    - Proceso de selección de metabolitos (o componentes) discriminantes en los

clasificadores posibles

- Comparación de modelos discriminantes de GRADE, INMUNO y GENOMICS.(Aquí hay que tener en cuenta que son distintas subpoblaciones (150, 150, 60)

BLOQUE 3: Propuesta de modelos generales, mixtos o generalizables posibles

Planteamiento de posibles modelos que incluyan otras variables covariables y efectos fijos o aleatorios observados

ANEXO. Tabla complementaria, **no necesaria para el desarrollo básico solicitado del ejercicio**. Describe cómo se construyen las categorías de CLASIFICACIÓN INMUNOHISTOQUÍMICA

Molecular typing of breast cancer based on common immnuohistochemical markers (Adb El-Rehim et al., 2005; Goldhirsch et al., 2011)

Molecular intrinsic subtype	Clinico-pathological definition	ER	PR	HER2	Ki67	Basal Markers*
Luminal A	**Luminal A**	+	+ or -	-	Low	-
Luminal B	**Luminal B (HER2 negative)**	+	+ or -	-	High	-
Luminal B	**Luminal B (HER2 positive)**	+	+ or -	Overexpressed	Low or High	-
HER2	**HER2 positive (non-luminal)**	-	-	Overexpressed	Usually High	+/-
Basal	**Triple negative (ductal)**	-	-	-	Usually High	+

* CK5/6 or CK14