***Análisis de un conjunto de datos de origen biológico mediante técnicas de machine learning supervisadas y no supervisadas***

**EQUIPO 2 LOTE 10**

Gálvez Robleño, Carlos

Rayo Morales, Mario Andrés

Benítez Fernández, Sergio

Salas Ramos, Maria

Encarnación Alejandro, Freyda Luisa

**Preguntas sobre las actividades:**

1. Procesamiento de los datos:

¿Qué método habéis escogido para llevar a cabo la imputación de los datos? Razonad vuestra respuesta. (0,3 puntos).

El método seleccionado para la imputación de los datos es el siguiente:

* **Imputación por la mediana** para las variables numéricas, que ayuda a evitar distorsiones causadas por valores extremos.
* **Imputación por la moda** para las variables categóricas.

¿Habéis llevado a cabo algún otro tipo de procesamiento? Razonad vuestra respuesta.

Sí, se han llevado a cabo varios otros procesos de procesamiento:

* **Eliminación de filas no relevantes** (como "sample\_0").
* **Escalado de las variables numéricas** para garantizar que todas tengan la misma importancia en los modelos.
* **Selección de características** utilizando Lasso para reducir la dimensionalidad y seleccionar las variables más relevantes.
* **División de los datos** en entrenamiento y prueba para evaluar el rendimiento de los modelos.
* **Entrenamiento y evaluación de modelos supervisados**.
* **Evaluación de modelos no supervisados.**

1. Métodos no supervisados:

¿Cuál es el motivo por el cual habéis seleccionado estas técnicas de reducción de dimensionalidad?

¿Cuál es el motivo por el cual habéis seleccionado estas técnicas de clusterización?

En ambos casos, ¿qué aspectos positivos y negativos tienen cada una?

|  |  |  |
| --- | --- | --- |
| **MÉTODOS** | **PROS** | **CONTRAS** |
| **ISOMAP** | Reduce dimensionalidad no lineal. | Costoso computacionalmente, sensible. |
| **T-SNE** | Bueno para visualización, captura relaciones locales. | Lento, difícil de interpretar, no adecuado para clusters grandes |
| **K-MEANS** | Simple, eficiente, escalable, fácil de entender | Requiere predefinir k, sensible a inicialización y valores atípicos |
| **WARD** | No requiere predefinir k, maneja clusters irregulares | Computacionalmente costoso, sensible al ruido y menos escalable |

En el caso de la clusterización, ¿podéis afirmar con certeza que los clústeres generados son los mejores posibles? Razonad vuestra respuesta.

1. Métodos supervisados:

¿Cuál es el motivo por el cual habéis seleccionado ambas técnicas de aprendizaje supervisado? ¿Cuál ha dado mejores resultados a la hora de clasificar las muestras? Razonad vuestra respuesta.

¿Habéis considerado oportuno implementar algún método de reducción de dimensionalidad para procesar los datos antes de implementarlos en dichas técnicas? ¿Por qué?

¿Qué aspectos positivos y negativos tienen cada una de las técnicas que habéis escogido?

* ***Técnica: RDA***

RDA es adecuado cuando tienes un número grande de genes en relación con el número de muestras, y cuando los datos pueden tener multicolinealidad o distribución no normal.

**Aspectos positivos:** Maneja bien la alta dimensionalidad y la colinealidad, previene sobre ajuste.

**Aspectos negativos:** Requiere ajuste de hiperparámetros, interpretación compleja.

* ***Técnica: LDA***

LDA puede ser útil si tus clases están bien separadas y los datos siguen distribuciones normales. Es más sencillo de interpretar y rápido de ejecutar, pero limitado si los datos son complejos.

**Aspectos positivos:** Simple y fácil de interpretar, eficiente computacionalmente, buen rendimiento con separaciones lineales.

**Aspectos negativos:** Supone normalidad y covarianza homogénea, limitado en separaciones no lineales.

* ***Técnica: SVM Lineal***

SVM lineal es útil si tus datos tienen muchas características (genes) y es probable que no sean linealmente separables, pero puede ser más costoso computacionalmente y más difícil de interpretar.

**Aspectos positivos:** Bueno para alta dimensionalidad, robusto frente a sobreajuste, puede manejar bien problemas no lineales.

**Aspectos negativos:** Requiere ajuste de parámetros, sensible al ruido, computacionalmente costoso.

1. De estas cuatro opciones, ¿qué tipo de arquitectura de *deep learning* sería la más adecuada para procesar datos de expresión génica? Razonad vuestra respuesta.

**a) Red de perceptrones *(multiperceptron layers).***

b) Redes convolucionales.

c) Redes recurrentes.

d) Redes de grafos.

Las redes de perceptrones multicapa (MLP, por sus siglas en inglés) son una de las arquitecturas de redes neuronales más comunes y efectivas para tareas de clasificación en datos tabulares, como los datos de expresión génica. A continuación, te explico por qué esta arquitectura es la más adecuada:

**Estructura tabular de los datos:** Los datos de expresión génica generalmente tienen una estructura tabular, donde cada fila representa una muestra (paciente, célula, etc.) y cada columna representa la expresión de un gen. Este tipo de datos es ideal para redes de perceptrones multicapa, ya que estas redes están diseñadas para trabajar con entradas estructuradas de esta manera (vectores de características).

**Capacidad de modelar interacciones complejas:** Las redes MLP son muy efectivas para capturar relaciones no lineales entre las variables, lo cual es crucial en el análisis de expresión génica. Los genes pueden interactuar de manera compleja, y un MLP tiene la capacidad de aprender esas interacciones sin la necesidad de que el usuario las especifique explícitamente.

**Flexibilidad:** Las redes MLP tienen una gran flexibilidad en términos de arquitectura (número de capas, número de neuronas por capa) y se pueden adaptar fácilmente a diferentes problemas. En el análisis de expresión génica, donde los conjuntos de

datos pueden variar mucho en tamaño y complejidad, las redes MLP pueden ajustarse para modelar de manera efectiva.

**Conclusión:**

Las redes de perceptrones multicapa son la opción más adecuada para procesar datos de expresión génica, ya que están diseñadas para trabajar con datos tabulares y tienen la capacidad de modelar interacciones complejas entre las variables (genes), lo cual es fundamental en este tipo de análisis.