Informe trabajo final programación
Susana Del Toro, Valentina Sánchez
Ingeniería biomédica
Universidad EIA

Introducción

El objetivo de este proyecto es evaluar y comparar dos enfoques de aprendizaje automático supervisado: Máquinas de Vectores de Soporte (SVM) y Clasificadores de Árboles de Decisión (DT). Utilizando el conjunto de datos *HCV-Egy-Data.csv*, buscamos determinar qué modelo es más adecuado para la clasificación de hepatitis C en pacientes egipcios, considerando métricas de rendimiento, optimización y curvas de aprendizaje.

Este estudio es relevante para identificar patrones en datos médicos y mejorar el diagnóstico temprano mediante herramientas computacionales.

Metodología

Base de Datos

- **Fuente:** *HCV-Egy-Data.csv*.
- **Características:** Incluye datos demográficos, clínicos y biomarcadores relevantes.
- **Preprocesamiento:** Se aplicaron técnicas para manejar valores faltantes y normalizar características según fuera necesario.

Modelos Utilizados

Máquina de Vectores de Soporte (SVM):

- Diseñada para problemas con límites de decisión no lineales.
- Implementación con kernel RBF (función de base radial) para capturar relaciones complejas en datos de alta dimensión.
- Optimización mediante GridSearchCV para ajustar parámetros como C (regularización), gamma y tipo de kernel.

Clasificador de Árbol de Decisión (DT):

- Modelo jerárquico que divide datos según las características más informativas.
- Fácil de interpretar y rápido, pero susceptible a sobreajuste si no se controla la profundidad.
- Optimización mediante GridSearchCV para parámetros como max_depth, min_samples_split y min_samples_leaf.

Resultados

1. **Desempeño del SVM:**
    - Mayor precisión, especialmente en datos complejos o de alta dimensión.

    - Tiempo de entrenamiento más largo debido al cálculo de kernels.
2. **Desempeño del DT:**
    - Modelos rápidos y fáciles de interpretar.
    - Posible sobreajuste si no se optimiza correctamente la profundidad del árbol.

Curvas de Aprendizaje

- Las curvas del SVM ayudan a ajustar parámetros y detectar desajustes.
- Las curvas del DT evidencian sobreajuste o subajuste, dependiendo de los datos de entrenamiento.

Discusión

- **SVM:** Ideal para problemas complejos donde se requiere alta precisión y la interpretación no es prioritaria.
- **DT:** Más adecuado para situaciones donde se necesita interpretabilidad y un análisis rápido, pero puede requerir regularización para evitar el sobreajuste.
- Impacto en el diagnóstico médico: ambos modelos son útiles, pero su aplicabilidad depende de los recursos computacionales y la necesidad de interpretabilidad en los resultados.

Conclusiones

- **Resultados clave:** El SVM tiende a ser más preciso en general, mientras que el DT es más rápido y comprensible.
- **Aplicaciones:** SVM es recomendable para datos de alta dimensión; DT es más adecuado cuando la transparencia del modelo es crucial.
- **Futuro:** Se sugiere explorar técnicas como ensambles (por ejemplo, Random Forest) o métodos de ajuste de hiperparámetros más avanzados.