Informe Final - Trabajo Programación 2025-1

1. Definición del problema

Se plantea un problema de clasificación binaria donde se desea predecir una variable de clase a partir de 15 atributos numéricos. El objetivo es identificar patrones que permitan anticipar la clase con base en los datos disponibles.

2. Descripción del dataset

El dataset contiene 5000 instancias y 15 atributos. Se introdujeron valores faltantes de forma aleatoria en el 5% de los datos para cumplir con los requerimientos del proyecto.

3. Análisis exploratorio

Se realizaron estadísticas descriptivas, análisis de valores faltantes y visualización de la correlación entre atributos. Esto permitió identificar patrones iniciales útiles para el entrenamiento de los modelos.

4. Preprocesamiento

Se imputaron los valores faltantes usando la media y se estandarizaron los datos usando `StandardScaler`. La variable objetivo fue separada del resto del conjunto.

5. Modelado y ajuste de hiperparámetros

Se entrenaron dos modelos:

- Random Forest con búsqueda de hiperparámetros usando GridSearchCV.

- SVM con búsqueda de hiperparámetros para C y kernel.

Ambos modelos fueron evaluados usando matriz de confusión y curva de aprendizaje.

6. Evaluación y diagnóstico

Se observaron buenos resultados con ambos modelos, aunque Random Forest mostró una ligera ventaja en precisión. Las curvas de aprendizaje indican un ajuste adecuado sin presencia significativa de overfitting o underfitting.

7. Comparación de modelos

Se compararon las métricas de desempeño de ambos modelos. Random Forest superó a SVM en precisión y recall. Se recomienda continuar con Random Forest para aplicaciones prácticas por su rendimiento y facilidad de interpretación.

8. Conclusiones

Este trabajo permitió aplicar todo el proceso de ciencia de datos, desde la exploración hasta la comparación de modelos. El enfoque propuesto demostró ser efectivo para resolver un problema predictivo con datos simulados.