PORTAFOLIOS DE MACHINE LEARNING

ESTUDIO DE CASO: PREDICCIÓN DE ENFERMEDAD DEL CORAZÓN

se presenta aquí un caso que versa sobre la predicción de enfermedad del corazón, a partir de una serie de predictores.

Se cuenta con 4 bases de datos referentes al diagnóstico de enfermedad del corazón, provenientes de:

1. Cleveland Clinic Foundation
2. Hungarian Institute of Cardiology, Budapest
3. V.A. Medical Center, Long Beach, CA
4. University Hospital, Zurich, Switzerland

Cada base de datos tiene el mismo formato de los ejemplos. Las bases tienen 76 atributos en total. En la bibliografía de UCI aparecen otros trabajos que, hasta el año 1989, sólo habían utilizado 14 de estos atributos.

Se desea en esta instancia seleccionar ***todos los atributos que sean significativos a efectos de generar modelos de ML que permitan hacer las mejores predicciones.***

La descripción (metadata) de las bases de datos se encuentra en el archivo **"heart-disease.names"**.

Las bases de datos provistas son:

1. **"cleveland.data"**
2. **"hungarian.data"**
3. **"long-beach-va.data"**
4. **"switzerland.data"**

El objetivo es, en este caso, predecir la **existencia de enfermedad del corazón**.

Líneas generales sobre los temas a considerar

1- Demostrar conocimiento organizacional o de contexto (conocimiento documentado del problema) – estudio del estado de la cuestión (investigar y documentar proyectos o estudios existentes sobe el tema en particular y el área en general)

2- Demostrar conocimiento detallado de los conjuntos de datos correspondientes

3- Demostrar capacidad en la preparación previa de los datos
   a. Integración de diferentes bases de datos
   b. Selección de features / atributos más relevantes para la predicción (no menos de 20)
   c. Análisis de estadísticas de los datos, detección de outliers, etc.

d. Importación, integración y preparación de los datos en RapidMiner (se deben tomar las 4 bases de datos como entrada al RM)
e. Manejo de los datos en Python y librerías asociadas

4- Algoritmos y modelos
   a. Selección justificada de método de aprendizaje automático para generar el modelo (entre los vistos hasta ahora)
   b. Selección de técnica para entrenamiento / validación , implementación en RapidMiner
   c. Generación del modelo en RapidMiner, con discusión de parámetros utilizados
   d. Análisis de la performance del modelo generado
   e. Realizar predicciones sobre un conjunto de ejemplos “no vistos” (describir y justificar la forma de separar este conjunto)
   f. Realizar análisis de utilización de estas predicciones, por medio de algunos ejemplos ilustrativos
   g. Realizar todas las etapas de preparación de datos, entrenamiento de algoritmos, estimación de indicadores de resultado (exactitud, precisión, recall, etc etc) en programas Python utilizando las librerías “pandas” y “scikit-learn”

5- Conclusiones generales sobre el caso abordado y la viabilidad u oportunidad de aplicación de técnicas de machine learning en el mismo.