Práctica final 2023/24

Título

Estudio sobre readmisión de pacientes con diabetes

Tareas

- La práctica se puede realizar de forma **individual** o en **parejas**;
- Descargar el conjunto de datos y revisar sus características/atributos (columnas);
- Probar diferentes modelos predictivos y ajustar sus parámetros sobre el conjunto de datos para obtener las mejores predicciones posibles;
- Comparar los resultados obtenidos;
- Crear un Google Colab con los pasos realizados, modelos probados con sus ajustes, resultados obtenidos y resúmenes organizados por niveles (ver *criterios de evaluación*).

Objetivo

- Usar la herramienta el lenguaje `Python` (`NumPy, Pandas, scikit-learn`) para construir modelos predictivos;
- Buscar y seleccionar modelos con la mayor precisión, utilizando como medida el *área bajo la Curva ROC* (ROC Área) mediante la estrategia de validación cruzada de 10 particiones (10-CV, Cross Validation).
- La variable a predecir es: **readmitted** (ubicada en última columna del fichero)

Nota: Dado que el conjunto de datos está desequilibrado, se recomienda generar o eliminar muestras para mejorar el aprendizaje de los modelos y evitar predicciones triviales.

Contenidos

Estudio sobre la **readmisión de pacientes con diabetes** (factores que influyen en el reingreso del paciente).

Se puede descargar en el siguiente enlace
https://archive.ics.uci.edu/dataset/296/diabetes+130-us+hospitals+for+years+1999-2008

- **Número de instancias:** 100.000
- **Número de atributos** (características): 47

Desarrollo

- Revisar cada característica (Ej. numérica/categórica, valores desconocidos, etc.), calcular el desequilibrio de la variable objetivo ('readmitted') para determinar el porcentaje de aciertos de un sistema trivial.
- Utilizar el clasificador base `sklearn.dummy.DummyClassifier()` como referencia básica para obtener el valor del área bajo la curva ROC (AUC) y mejorarlo en las siguientes pruebas.
- Probar diferentes clasificadores con el objetivo de obtener el mayor AUC posible, ajustando parámetros manualmente y evaluando diferentes algoritmos.
- Explorar la posibilidad de selección automática de atributos con diferentes algoritmos.
- Comparar los resultados (10-CV) usando la media aritmética, o con la prueba de contraste de hipótesis de Wilcoxon (signed-rank test).
- Tener en cuenta el elevado número de instancias del conjunto de datos y su impacto en el tiempo de procesamiento.

Criterios de evaluación

Nivel básico (5 puntos)

- Adaptar el tipo de datos para una clasificación correcta (Ej.: detectar qué variables aparentemente numéricas no lo son por tener cadenas de texto, eliminar dichas muestras y convertir las columnas en numéricas).
- Aplicar al menos 2 clasificadores distintos al base (`DummyClassifier`).
- Cambiar parámetros manualmente para mejorar el AUC;
- Redacción clara de la memoria explicando las tareas realizadas y los resultados obtenidos.

Nivel medio (3 puntos)

- Probar algoritmos de selección automática de atributos.
- Aplicar al menos un total de 5 clasificadores.
- Utilizar algoritmos de reducción de dimensionalidad antes de la clasificación.
- Realizar comparativas de los resultados con contraste de hipótesis (Ej. Wilcoxon signed-rank test).
- Redacción de la memoria adecuada para este nivel con explicaciones claras y detalladas de los apartados, ofreciendo resúmenes de los mejores resultados.

Nivel avanzado (2 puntos)

- Buscar parámetros óptimos de forma automática (Ej. grid search) para algunos clasificadores para intentar mejorar el AUC (nested cross-validation);
- Aplicar razonamientos o técnicas relacionadas con la práctica no estudiadas directamente en clase pero útiles para el estudio del conjunto de datos.
- Usar algún clasificador avanzado (no pertenecientes a scikit-learn), o redes neuronales más avanzadas que un MLP especialmente diseñadas para datos tabulares.

- Redacción de un informe excepcional, bien estructurado según niveles de evaluación, con claridad, bien redactado, con resúmenes de los resultados correctamente justificados.

Entrega

- La práctica se podrá entregar hasta el **miércoles, 13 de diciembre a las 23:59h**;
- **Completar el siguiente formulario** con el enlace de **Google Colab compartido en abierto** (pasos: Compartir / Acceso general / Cualquier persona con el enlace) para que se pueda revisar, junto a una nota solicitada y su breve justificación correspondiente.