Práctica 4 sobre Inteligencia Artificial Explicable

Instalación y aprendizaje

En esta parte de la práctica vamos a aprender a usar los explainers SHAP, ALE, LIME y DiCE sobre datos tabulares. Revisa e instala las siguientes librerías:

- SHAP

Instalación en anaconda: *pip install shap* (si no os funciona podéis incluir *!pip install shap* en la primera celda del notebook).

Revisa la librería donde verás algunos ejemplos de aplicar SHAP (Kernel) a varios modelos de IA: https://shap.readthedocs.io/en/latest/example_notebooks/tabular_examples/model_agnostic/Iris%20classification%20with%20scikit-learn.html

- ALE

Instalación en anaconda: *pip install alibi[all]*

Revisa la librería, donde se puede observar cómo implementar un ejemplo: https://docs.seldon.io/projects/alibi/en/latest/methods/ALE.html

- LIME

Instalación en anaconda: *pip install lime* en la consola de anaconda o *!pip install lime* en una celda del notebook

La documentación de la librería está detallada en GitHub (https://github.com/marcotcr/lime). Revísala para ver ejemplos y tutoriales.

- DiCE

Instalación en anaconda: pip install dice-ml en la consola de anaconda o *!pip install dice-ml* en una celda del notebook

La documentación de la librería se encuentra en esta página web: https://interpret.ml/DiCE/index.html En la primera página se muestra cómo importar la librería y un pequeño ejemplo.

Práctica

Debéis escoger una de las configuraciones del perceptrón multicapa que estudiasteis en la parte A de la práctica 3. Vamos a aplicar los siguientes modelos de XAI para explicar el comportamiento de ese perceptrón.

1. Usaremos SHAP para explicar **una única predicción del modelo para una única instancia**. La explicación mostrará la importancia de cada característica en la predicción. Aunque se pueden generar muchos gráficos con SHAP, utilizaremos el gráfico de **waterfall,** que es el más común (para visualizar los gráficos, es posible que necesitéis ejecutar *shap.initjs()* en una celda del notebook). En este enlace tienes un ejemplo de cómo aplicar SHAP y generar ese gráfico:

https://shap.readthedocs.io/en/latest/example_notebooks/overviews/An%20introduction%20to%20explainable%20AI%20with%20Shapley%20values.html

   a. Muestra la predicción obtenida con el modelo para esa instancia y el valor real que tenemos en el conjunto de datos.
   b. Describe la gráfica resultante obtenida con SHAP. ¿Concuerda la explicación de SHAP con la predicción del modelo y/o el valor real?
   c. Determina cómo afectan los atributos de colesterol y presión para esa predicción. ¿Cuáles de ellos afectan de forma positiva o negativa a tener un problema cardiovascular? ¿Cuál es el atributo que afecta más?

2. Usaremos ALE para generar **una explicación global** (muestra el comportamiento general del modelo). Replica lo visto en el ejemplo de la librería para usar ALE sobre vuestro perceptrón multicapa. Determina cómo afectan los atributos de colesterol y presión sobre el comportamiento del modelo.
3. Usaremos LIME para explicar **una única predicción del modelo para una única instancia (la misma instancia que ya habéis explicado con SHAP)**. La explicación mostrará la importancia de cada característica en la predicción. Revisad el tutorial (https://marcotcr.github.io/lime/tutorials/Tutorial%20-%20continuous%20and%20categorical%20features.html) donde se muestran varios ejemplos de implementación de LIME y su interpretación. Vamos a analizar la explicación de la misma forma que analizamos SHAP:
   a. Describe la gráfica resultante obtenida con LIME. ¿Concuerda la explicación de LIME con la predicción del modelo y/o el valor real?
   b. Determina cómo afectan los atributos de colesterol y presión para esa predicción. ¿Cuáles de ellos afectan a tener un problema cardiovascular? ¿Cuál es el atributo que afecta más?
4. Usaremos DiCE para **obtener contraejemplos de la misma instancia que hemos estudiado con SHAP y LIME. Obtendremos 3 contraejemplos**. En la documentación (aquí: https://interpret.ml/DiCE/notebooks/DiCE_getting_started.html) se puede ver un ejemplo de implementación, revísalo para replicarlo en el modelo de perceptrón multicapa aplicado sobre la detección de problemas cardiovasculares.
   a. Describe los contraejemplos obtenidos con DiCE.
   b. Determina cómo afectan los atributos de colesterol y presión para esa predicción. ¿Cuál es el atributo que afecta más en la predicción?

Determina unas **conclusiones finales**. ¿Concuerdan las diferentes explicaciones entre ellas? ¿Qué diferencias hay entre los diferentes tipos de explicaciones ofrecidas y cuáles son más adecuadas para explicar unos tipos de problemas u otros?

Normas de Entrega

La práctica debe entregarse utilizando el mecanismo de entregas del campus virtual. Se entregarán uno o dos notebooks (uno por cada parte), como prefiráis, con el código en Python explicando el análisis de resultados y el código realizado. Sin las explicaciones del código realizado y el análisis de los resultados obtenidos, la práctica será no apta.