Tema 5
Sistemas de Información Sanitaria
Machine Learning interpretable

Ingeniería en Tecnologías de la Telecomunicación

Escuela de Ingeniería de Fuenlabrada

Universidad Rey Juan Carlos

Introducción

- *Interpretable machine learning (IML), explainable machine learning* o *explainable artificial intelligence* (XAI) se define son un conjunto de métodos y técnicas por los cuales un usuario o experto que usa *machine learning* puede entender la lógica que subyace a las decisiones y resultados del modelo.
- Importante para explicar a los clínicos por qué el modelo está tomando una decisión.

Introducción

- Algunos métodos de *machine learning* son intrínsicamente interpretables (*white-box models*):
  - Regresión logística
  - *Decision trees*
- Otros métodos no son intrínsicamente interpretables (*black-box models*):
  - *Random forest*
  - *Xgboost*
  - Redes neuronales

Introducción

- Tipos de métodos:
  - *Model-specific* vs. *Model-agnostic*
    - Las técnicas de interpretabilidad de específicas se limitan a tipos concretos de modelos, mientras que las técnicas agnósticas pueden aplicarse a cualquier modelo (se denominan también técnicas post hoc).
  - Local vs. Global
    - Los métodos de interpretabilidad locales se refieren a métodos que explican una única predicción, mientras que los métodos globales se refieren a métodos que explican el modelo global.

Introducción

- Ejemplos de métodos:
  - *Model-specific:*
    - Globales:
      - Ver la importancia de cada variable en un árbol de decisión
      - Ver los pesos de cada variable en una regresión logística
  - Model-agnostic:
    - Globales:
      - Partial Dependence Plots (PDP)
      - Individual Conditional Expectation (ICE) plots
    - Locales:
      - SHapley Additive exPlanations (SHAP)
      - Local Interpretable Model-agnostic Explanations (LIME)

Trabajo

- El trabajo consistirá en realizar un documento de una extensión máxima de 5 páginas en el que se explicará qué es el *machine learning* interpretable (en inglés llamado *interpretable machine learning* [IML], *explainable machine learning* o *explainable artificial intelligence* [XAI]), su importancia para en el ámbito de la salud y biomédico cuando se usan métodos de *machine learning*, y algún ejemplo de aplicación. Debéis tener una sección del método SHAP, ya que se hará una práctica de este método.
  - Se deberá citar de manera rigurosa las fuentes usadas para la realización del trabajo.

Bibliografía

- Christoph Molnar. Interpretable Machine Learning: A Guide for Making Black Box Models Explainable
  - https://christophm.github.io/interpretable-ml-book/
- Patrick Hall, Navdeep Gill. An Introduction to Machine Learning Interpretability: An Applied Perspective on Fairness, Accountability, Transparency, and Explainable AI
  - https://h2o.ai/content/dam/h2o/en/marketing/documents/2019/08/An-Introduction-to-Machine-Learning-Interpretability-Second-Edition.pdf
- Przemyslaw Biecek and Tomasz Burzykowski. Explanatory Model Analysis: Explore, Explain, and Examine Predictive Models
  - https://ema.drwhy.ai/

Bibliografía

- Conor O'Sullivan. What is Explainable AI (XAI)? An introduction to XAI — the field aimed at making machine learning models understandable to humans
  - https://medium.com/data-science/what-is-interpretable-machine-learning-2d217b62185a
- Sheng-Chieh Lu et al. On the importance of interpretable machine learning predictions to inform clinical decision making in oncology
  - https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10013157/