Unidad	Unidad 2 Tipos de Explicabilidad de Algoritmos
Entrega	Método de entrega de la actividad en el campus virtual

1. Enunciado

En esta actividad deberéis seleccionar un modelo de aprendizaje automático más complejo, no interpretable que se pueda aplicar al mismo dataset que seleccionaron en la tarea 1. Como ya hicieron el análisis exploratorio del dataset anteriormente, no hace falta repetirlo. Por tanto, solo tenéis que entrenar el modelo seleccionado con el mismo dataset, comparar el rendimiento obtenido con ambos modelos y explorar la interpretabilidad del modelo.

El objetivo principal tiene dos partes:

1. Comparar el rendimiento obtenido con modelos interpretables (más sencillos) y modelos poco interpretables (más complejos, pero más potentes), incluyendo el tiempo de cómputo consumido con cada modelo.
2. Comparar las posibilidades de interpretabilidad de ambos modelos, aplicando métodos agnósticos al nuevo modelo.

Además, la realización de esta actividad tiene los siguientes objetivos secundarios:

- Entrenar y evaluar el modelo de aprendizaje automático no interpretable seleccionado
- Aplicar métricas de rendimiento estándar y
- Evaluar la interpretabilidad del modelo y de sus predicciones, comprendiendo la importancia de las métricas de interpretabilidad y los métodos agnósticos al modelo.
- Documentar y comunicar de manera efectiva sus procesos y hallazgos, presentando visualizaciones, análisis y resultados detallados y
- Participar en discusiones constructivas para recibir y proporcionar retroalimentación.

2. Pasos para realizar la actividad

- Selección del modelo de aprendizaje automático no interpretable apropiado para el análisis del conjunto de datos.
- Entrenar el modelo de aprendizaje automático con el conjunto de datos preparado.
    - Evalúa el rendimiento inicial del modelo utilizando métricas estándar.
- Evaluación de la interpretabilidad aplicando métodos modelo-agnósticos.
- Documentar todo el proceso, tanto la evaluación del rendimiento inicial del modelo utilizando métricas estándar como la evaluación de la interpretabilidad del mismo. Incluye visualizaciones, resultados y análisis detallados.

- Preparar una presentación para compartir tus hallazgos con la clase. La presentación debe destacar las técnicas utilizadas, los resultados obtenidos y las posibles mejoras en la interpretabilidad del modelo.
- Discusión y retroalimentación: Presenta tu trabajo a la clase y participa en una discusión abierta sobre los desafíos y hallazgos del análisis de interpretabilidad.

3. Detalles de la entrega

Documentos a entregar a través de la plataforma:

1. Informe sobre la selección del modelo de aprendizaje automatizado no interpretable, explicando los diferentes pasos realizados para su entrenamiento y la evaluación de sus métricas de desempeño y de interpretabilidad.
2. Código del programa utilizado para entrenar y evaluar el modelo en formato (.py, .ipynb, etc.) que facilite su ejecución mediante el IDE adecuado
3. Presentación para compartir sus hallazgos con la clase, destacando las técnicas utilizadas, resultados obtenidos y posibles mejoras en la interpretabilidad del modelo

Subir de forma individual los documentos a la actividad 1 del campus virtual.

Recuerda (11pt negrita)

Al ser un ejercicio puntuable, se evaluarán los aspectos ya señalados. Además, se puntuarán muy firmemente los siguientes:

- La originalidad: es decir, las copias de Internet, de un compañero o de otra bibliografía no serán válidas.
- La simplicidad: enunciados claros y soluciones fáciles de seguir.
- La solución: no es suficiente con unos buenos ejercicios si estos están mal resueltos.
- La planificación: entregas a tiempo y consulta con el profesor la evolución de la tarea.

4. Anexo

Criterios para los conjuntos de datos utilizados:

a) Deben ser conjuntos de datos provenientes de la realidad cotidiana (investigaciones, encuestas) tomados de páginas de los gobiernos con estadísticas, de competencias de Kaggle u otras similares, repositorios como UCI, etc.)
b) No se aceptarán conjuntos de datos obtenidos de forma artificial mediante código creado para generarlos como se acostumbra para hacer demostraciones en las páginas web y ejemplos de cursos.
c) Los conjuntos de datos deben tener más de 1000 muestras.
d) **No** se aceptarán para las tareas ninguno de los siguientes conjuntos de datos o versiones similares
   - scikit-learn

- MNIST
- California Housing
- Iris flower
- Breast cancer
- Diabetes
- cats_vs_dogs
- Heart disease
- Titanic (cualquier versión de este dataset)
    - https://raw.githubusercontent.com/datasciencedojo/datasets/master/titanic.csv
    - https://www.kaggle.com/competitions/titanic/data
    - https://www.kaggle.com/datasets/yasserh/titanic-dataset

Bibliografía:

1. Molnar, C. (2025). Interpretable Machine Learning: A Guide for Making Black Box Models Explainable (3rd ed.), christophm.github.io/interpretable-ml-book/
2. Molnar, C. (2021) Aprendizaje automático interpretable: Una guía para hacer que los modelos de caja negra sean explicables. https://fedefliguer.github.io/AAI/