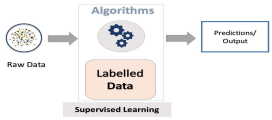


UNIDAD 2 - PRACTICA 5
ARBOLES DE CLASIFICACION en PYTHON

U2.T6.1.-

La página web

https://eldave93.github.io/Machine-Learning-in-Python-20-21/Week_09_Trees/1_Decision_Trees.html

fue comentada en la sesión del 16 de diciembre de 2023 y figura en el módulo de CANVAS de dicho día. Es un ejemplo de código en PYTHON para construir árboles de clasificación y calcular medidas de exactitud, utilizando el conjunto de los 342 pingüinos de 3 especies diferentes y de 3 islas del archipiélago Palmer, en la Antárctica. El segundo ejemplo con los datos sobre "Breast Cancer Wisconsin" es muy interesante al incluir los análisis de pre-poda y post-poda.

Ambos ejemplos se deben ampliar con la selección de un conjunto de entrenamiento (TR) y un conjunto de test (TS), obteniendo la matriz de confusión y otras medidas de error en cada uno de esos conjuntos.

Mediante Jupyter Notebook, cuyo nombre será 2024-Enero-U2.T6-1-PING-Apellido1-Apellido2-Nombre, proceder a

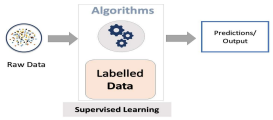
a) Añadir comentarios al citado código (sólo la parte de los pingüinos que acaba en la celda [26]), indicando lo que realiza cada celda o bloque. Comentar si es posible programar alguna celda de manera más eficiente a la que muestra la sintaxis que está en la web anterior.

b) Interpretar las salidas gráficas que se generan al ejecutar cada una de las celdas.

U2.T6.2.-

La predicción de una situación de salud como la posible presencia de un virus en una persona, o encontrarnos con una persona ante la que dudamos si hay un ataque al corazón o no, es una aplicación de los modelos de ML de aprendizaje supervisado.

Siguiendo https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction, las enfermedades cardiovasculares (CVD por sus siglas en inglés) son la primera causa de muerte en el mundo: se calcula que cada año se cobran 17,9 millones de vidas, lo que representa el 31% de todas las muertes en el mundo. Cuatro de cada cinco muertes por CVD se deben a infartos de miocardio y accidentes



cerebrovasculares, y un tercio de estas muertes se producen prematuramente en personas menores de 70 años. La insuficiencia cardiaca es un evento común causado por las CVD y este conjunto de datos contiene 11 características que pueden utilizarse para predecir una posible enfermedad cardiaca.

Las personas con enfermedades cardiovasculares o que corren un alto riesgo cardiovascular (debido a la presencia de uno o más factores de riesgo como hipertensión, diabetes, hiperlipidemia o enfermedad ya establecida) necesitan una detección y gestión tempranas en las que un modelo de aprendizaje automático puede ser de gran ayuda.
(Traducción realizada con la versión gratuita del traductor DeepL.com)

La página web

https://www.kaggle.com/code/beingamit99/heart-disease-decision-tree

es bastante completa como ejemplo de la problemática anterior. Muestra todo el código en PYTHON.

Mediante Jupyter Notebook, cuyo nombre será 2024-Enero-U2.T6-2-CVD-Apellido1-Apellido2-Nombre, proceder a

- **a)** Añadir comentarios al citado código, indicando lo que realiza cada celda o bloque. Comentar si es posible programar alguna celda de manera más eficiente a la que muestra la sintaxis que está en la web anterior.
- **b)** Interpretar las salidas gráficas que se generan al ejecutar cada una de las celdas.
- **c)** Incorporar celdas adicionales que mejoren el código tanto en términos de ML, como de Estadística, como de Programación. Por ejemplo, incluir para algunas variables críticas en la problemática CVD, un gráfico del tipo Figura 3 (Penguins Data Pairplot) del documento citado en el ejercicio anterior U2.T6.1.