PRÁCTICA 1: APRENDIZAJE AUTOMÁTICO

PARTE A: análisis y procesamiento de un dataset

(Se necesita utilizar la librería Pandas para realizar la práctica). En esta parte usaremos el dataset que está en el campus virtual: smart_agriculture_bangladesh.csv. El dataset recopila datos, recogidos a través de sensores IoT, sobre el medio ambiente en distintas zonas agrícolas de Bangladesh. Contiene información sobre la temperatura, humedad, etc.

Sobre este dataset, hay que realizar las siguientes tareas:

1. Carga el dataset con la librería pandas. Describe cada atributo del dataset: defínelo describiendo el atributo, determina si es un atributo categórico (nominal, binario, ordinal), continuo o textual.
2. Elimina los duplicados y valores nulos (pista: ambas operaciones se realizan con una única línea de código cada una). ¿Cuántas filas se han borrado? ¿En qué beneficia la eliminación de estos valores? (Al eliminar fiilas, se debe usar después reset_index, que resetea los índices para evitar problemas en los ejercicios siguientes).
3. Antes de empezar, determina si hay algún atributo que no nos va a resultar útil y por qué. Eliminados si hay.
4. Atributos continuos:
    a. Calcula media, desviación típica, valores mínimos, máximos, etc. de los atributos numéricos. Describe estos valores para cada variable.
5. Atributos categóricos:
    a. Dibuja histogramas, diagramas de barras o de tartas para determinar las frecuencias de los valores de los atributos categóricos. Indica el número de valores distintos para cada atributo y el valor más frecuente para cada atributo. ¿Qué atributos están balanceados y cuáles no?
6. Determina si hay outliers. Fíjate en las gráficas y descripciones que has realizado antes.
7. Convierte los atributos categóricos en valores numéricos usando OneHotEncoder (o getdummies) y LabelEnconder. Observa las diferencias y discute cuál sería la mejor opción cuando el dataset sea usado en un modelo de IA.
    a. Relaciones entre atributos: dibuja diagramas de dispersión y calcula coeficientes de correlación. ¿Cuáles son los atributos que están más relacionados y qué podemos interpretar?
8. Normaliza y estandariza el dataset. Observa los resultados. Discute qué opción sería mejor usar: normalización o estandarización.

PARTE B: evaluación de modelos de AA

Instalación

Para instalar la librería que se necesita, vamos a instalar la librería Scikit-learn:

https://scikit-learn.org/ Para instalar Scikit-learn se necesita los siguientes requisitos previos:

- Python (>= 2.6 or >= 3.3),
- NumPy (>= 1.6.1),
- SciPy (>= 0.9).

En la siguiente web, se explica cómo instalar esta librería para cada sistema operativo:

https://scikit-learn.org/0.16/install.html

Por ejemplo, en el caso de Windows, se debe ejecutar lo siguiente desde la consola de comandos:

$> pip install -U scikit-learn

Puede variar dependiendo como tengáis configurado vuestro entorno de Python en vuestro editor/herramienta correspondiente.

Aprendizaje

Tómate tu tiempo para explorar a fondo la web de Scikit-learn: https://scikit-learn.org/

Mira cómo está organizada la web y qué grandes apartados de aprendizaje automático soporta. Entre en cada uno de ellos viendo toda su estructura y entra a leer aquellos aspectos que te llame más la atención.

En concreto, revisa los modelos de clasificación kNN (https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html) y RandomForest (https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html). Revisa los parámetros que se pueden configurar para ambos modelos.

Práctica

Sobre el dataset anterior, y partiendo del análisis ya realizado, vamos a resolver un problema de clasificación. Vamos a predecir el riesgo de plagas (variable pest_risk) usando dos modelos de clasificación distintos: kNN y Random Forest.

Tareas a realizar:

1. Usad random_state para crear la validación cruzada y entrenar Random Forest (kNN no lo necesita porque es determinista). Esta variable se la vamos a pasar a los modelos para que cada vez que se ejecute el código salgan los mismos resultados. De esta forma os aseguráis de que los resultados que os salgan serán los mismos que me salgan a mí al ejecutar el código. Si no hacéis esto, vuestro análisis puede no tener ningún sentido en mi ejecución y os arriesgáis al no apto.
2. Para cada modelo (kNN y RandomForest), decidid si vais a usar OneHotEncoder o LabelEncoder para las variables categóricas. Justifica por qué.
3. Para cada modelo (kNN y RandomForest), decidid si vais a usar escalado, estandarización o ninguna de las dos. Justifica por qué.
4. Configura una validación cruzada de **5 particiones**. IMPORTANTE: los datasets tienen que tener las instancias mezcladas aleatoriamente.
5. Configura el modelo kNN usando 7 vecinos más cercanos y la métrica de distancia Euclídea. El resto de la configuración la dejaremos por defecto.
6. Configura el modelo Random Forest con n_estimators=100, max_depth=None, min_samples_split=3, min_samples_leaf=1. Explica en qué consiste cada parámetro y los valores que les hemos puesto.
7. Calcula los valores de las métricas de evaluación precisión, recall y F1, para ambos modelos y su matriz de confusión.
8. Analiza los resultados para cada modelo y cada métrica, y compáralos determinando cuál es el mejor modelo y por qué.

PARTE OPCIONAL (contribuye a 1 punto adicional en la nota final de la asignatura)

Usar randomized search y gridsearch para ambos modelos kNN y Random Forest. Determina los mejores parámetros encontrados. Discute los resultados de las métricas de evaluación precisión, recall, F1 y la matriz de confusión para cada modelo y compáralos entre ellos. Compara estos resultados con los resultados obtenidos en la parte B.

Normas de entrega

Se entregará 1 notebook en la entrega habilitada en el campus virtual. Recuerda que todas las celdas deben funcionar para que sean corregidas. **Muy importante**: contesta a las preguntas del enunciado y comenta el código y los resultados. Sin las respuestas a las preguntas y el análisis de los resultados, la práctica no conseguirá el apto.