Informe Técnico – Clasificación y Agrupamiento de Especies de Flores

1. Introducción

El objetivo de este estudio es explorar y comparar dos enfoques de Machine Learning aplicados al famoso dataset Iris, que contiene medidas de sépalos y pétalos de tres especies de flores: setosa, versicolor y virginica.

Se implementaron dos paradigmas:
1. Aprendizaje Supervisado: Entrenamiento de un modelo con etiquetas conocidas para predecir la especie de nuevas flores.
2. Aprendizaje No Supervisado: Agrupamiento de las flores por similitud sin usar etiquetas, buscando descubrir la estructura natural del dataset.

Estas dos aproximaciones permiten entender tanto el poder predictivo de los modelos como su capacidad para identificar patrones ocultos en los datos.

2. Metodología y Resultados – Aprendizaje Supervisado

Preparación de Datos:
- Se separaron las características (X) de la variable objetivo (y). Luego, se dividió el dataset en 80% entrenamiento y 20% prueba, asegurando que la proporción de especies se mantuviera (stratify=y).

Modelo: Árbol de Decisión:
- Se utilizó un DecisionTreeClassifier, un modelo interpretable que aprende reglas de decisión basadas en las características de cada flor.
- El modelo fue entrenado con los datos de entrenamiento (X_train, y_train).

Evaluación:
- Se realizaron predicciones sobre el conjunto de prueba (X_test) y se calculó la precisión (accuracy).
- Accuracy obtenida: 100%, indicando que el modelo aprendió correctamente las relaciones entre características y especies.

Discusión:
- El Árbol de Decisión separó efectivamente las especies gracias a la información clara y linealmente separable del dataset, especialmente para setosa.
- Ventaja: alta precisión y explicabilidad. Desventaja: requiere datos etiquetados y puede sobreajustarse.

3. Metodología y Resultados – Aprendizaje No Supervisado

Determinación del número de clústeres:
- Para aplicar K-Means, se utilizó el Método del Codo, graficando la inercia (suma de distancias al cuadrado) para distintos valores de K.
- El codo apareció en K=3, coincidiendo con el número real de especies.

Aplicación de K-Means:
- Se entrenó K-Means con K=3 sobre todas las características y se obtuvieron las etiquetas de clúster y los centroides.

Visualización:
- Scatter plot de petal length vs petal width, coloreando por clúster y mostrando centroides.
- Se observa que los clústeres separan mayormente las especies, con setosa claramente diferenciada.

Tabla de Contingencia:
- Comparación de clústeres con especies reales usando pd.crosstab.
- Algunos grupos muestran mezclas entre versicolor y virginica, indicando limitaciones del clustering.

4. Discusión Comparativa y Conclusión

Comparativa de métodos:
- Aprendizaje Supervisado: requiere etiquetas, alta precisión, interpretabilidad, predice especies con certeza.
- Aprendizaje No Supervisado: no requiere etiquetas, descubre patrones ocultos, puede mezclar grupos similares.

Conclusión:
- Supervisado es eficiente cuando se dispone de etiquetas confiables.
- No supervisado permite explorar la estructura intrínseca de los datos, aunque con limitaciones.
- Las especies del dataset Iris presentan diferencias claras, lo que facilita el aprendizaje supervisado.
- K-Means descubre clústeres naturales, pero no siempre coincide perfectamente con la realidad biológica.