Supervised vs. Unsupervised Learning

Notes for SHS • Data Science Topic 55 • Answer guide

Work from concepts to responsible interpretation. This worksheet accompanies two reproducible Python projects using the Iris data.

1. Define the two methods

Explain supervised and unsupervised learning in one or two sentences.

Explanation: Supervised models are fitted using examples with known target outcomes. Unsupervised methods investigate features without those target labels supplied.

2. Identify labels

A flower table has four measurements and a known species column. Which is the target?

Explanation: Species is the target label; flower dimensions are input features.

3. Classification or regression?

Predict the precise temperature in °C from recent measurements.

Explanation: Regression, because the intended output is a number on a meaningful continuous scale.

4. Classification or clustering?

Group unlabeled soil samples by similarities in their measured properties.

Explanation: Clustering, a typical unsupervised task that investigates groupings without supplied target categories.

5. Feature selection

Why could scaling matter when variables are measured in different units?

Explanation: Many distance-based methods, including K-Means, are affected by variable scales; standardization gives features comparable scale, though may not always be appropriate.

6. Train/test split

Why keep held-out observations away from supervised model fitting?

Explanation: Independent held-out labels help estimate generalization instead of measuring how well the model recalls training samples.

7. Accuracy

A classifier correctly identifies 34 of 38 flower samples. Calculate accuracy.

Explanation: 34/38 ≈ 0.8947, or 89.47%, for this specific split only.

8. Cluster identifiers

What does K-Means Cluster 0 mean? Is it a flower species?

Explanation: The cluster ID is an arbitrary group name assigned by the algorithm, not the botanical name or a proven natural species category.

9. Evaluation

Why cannot classification accuracy be compared directly with silhouette score?

Explanation: Accuracy compares predictions with true labels, whereas silhouette describes cohesion and separation of data clusters. They measure different objectives.

10. Code challenge

Run both supplied Python scripts. Change max_depth for the tree and n_clusters for K-Means. Describe changes without claiming a single best answer.

Explanation: Results depend on dataset and settings; record actual observations, judge with appropriate validation, and never tune using the final test set.

References: https://scikit-learn.org/stable/ & https://developers.google.com/machine-learning/crash-course