how.To → ASSESS ALGORITHM FIT

UNSUPERVISED	Pitfalls, techniques for clustering, dimension reduction, outlier by algorithm.
SUPERVISED	Pitfalls, evaluation methods, classify, consensus, assessment approaches.

ASSESS MODEL FIT ⇒ UNSUPERVISED

Evaluation methods encompass various aspects specific to each algorithm category.

- **Clustering methods** focus is on quantifying the formed clusters' quality, consistency, and discriminative ability.
- **Dimensionality reduction techniques** aim to assess information retention, preservation of data relationships, and visualization capabilities.
- **Outlier detection** gauges the accuracy and efficacy of identifying anomalies in the dataset.

UNSUPERVISED - Evaluation pitfalls

Mitigate pitfalls by employing a comprehensive evaluation, including cross-validation techniques, comparison to baselines or benchmarks, and consulting with domain expertise.

- **Overfitting:** Use evaluation methods with caution to avoid overfitting algorithms to specific evaluation datasets. Algorithms that perform exceptionally well on the evaluation data may need to generalize to new, unseen data.
- **Metric Limitations:** Each evaluation metric has its limitations and underlying assumptions. Understand each measured characteristic and select those responding to the features of the dataset and project analysis objectives. Data scientists often set up functions to run multiple tests and visually assess the outcomes to gain a comprehensive understanding.
- **Dataset Bias:** Dataset bias refers to situations where the dataset information does not represent an underlying population's actual characteristics. To help address dataset bias, ask informed questions based on thorough scientific research and knowledge of the measuring instrument, such as the choice of Likert scales, logarithmic transformations, and advanced data substitution methods like bootstrapping. Most datasets have a "data dictionary" detailing how each variable is measured.
- **Subjectivity in Interpretation:** Evaluation metrics involve subjective interpretation or require domain knowledge to comprehend their implications. Learn any contextual factors and subjectivity involved.

UNSUPERVISED - Techniques to discern prediction quality

A. Clustering techniques

- **Evaluate Cluster Quality:** Calculate metrics such as Silhouette Score, Davies-Bouldin Index, or Calinski-Harabasz Index to assess the quality of the clusters.
- **Visualize Clusters:** Plot the clusters in a 2D or 3D space using dimensionality reduction techniques like PCA or t-SNE for visual inspection.
- **Parameter Tuning:** Iterate over different values for the algorithm-specific parameters (e.g., number of clusters, distance metric) and evaluate the impact on cluster quality.
- **Compare with Ground Truth (if available):** If you can access ground truth labels, use metrics like Rand Index or Adjusted Rand Index to compare the predicted clusters with the true ones.

B. Dimensionality reduction techniques

- **Evaluate Data Representation:** Calculate metrics such as Explained Variance Ratio or Kullback-Leibler Divergence to assess how well the technique represents the original data.
- **Visualize Reduced Dimensions:** Plot the reduced dimensions (e.g., principal components, t-SNE embeddings) to examine the separation or clustering patterns visually.
- **Assess Information Retention:** Analyze the cumulative explained variance or other metrics to understand restrained information in the reduced dimensions.
- **Consider Downstream Performance:** If the dimensionality reduction is a precursor to another task, evaluate the performance of the downstream task using the reduced dimensions.

C. Outlier detection techniques

- **Evaluate Data Representation:** Calculate metrics such as Explained Variance Ratio or Kullback-Leibler Divergence to assess how well the technique represents the original data.
- **Visualize Reduced Dimensions:** Plot the reduced dimensions (e.g., principal components, t-SNE embeddings) to examine the separation or clustering patterns visually.
- **Assess Information Retention:** Analyze the cumulative explained variance or other metrics to understand restrained information in the reduced dimensions.

- **Consider Downstream Performance:** If the dimensionality reduction is a precursor to another task, evaluate the performance of the downstream task using the reduced dimensions.

Clustering techniques

Clustering and dimensionality reduction rely on ground truth labels and specific metrics to assess performance. The following is a small portfolio of available techniques, and you will use those with numbers in front of them to determine outcomes in your Iris flower dataset assignment.

D. Clustering algorithms (K-means, hierarchical clustering, DBSCAN)

- **Silhouette Score:** Measures the compactness and separation of clusters.
  - Use to evaluate the quality of clustering results by measuring the compactness and separation of the clusters in the Iris dataset.
- **Calinski-Harabasz Index:** Measures the ratio between intra-cluster and inter-cluster variance.
  - In the Iris dataset, it measures the ratio between the intra-cluster and inter-cluster variance, providing insights into the clustering quality.
- **Davies-Bouldin Index:** Evaluates the clustering quality based on intra-cluster and inter-cluster distance.
- **Rand Index:** Compares the similarity between predicted and true clusters.
- **Gap Statistic:** Measures the optimal number of clusters by comparing the within-cluster dispersion with that expected under a null reference distribution.
- **Elbow Method:** Plots the explained variance or distortion as a function of the number of clusters and identifies the "elbow" point where the rate of improvement diminishes significantly.
- **Hopkins Statistic:** Measures the clustering tendency or the presence of clusters in the data by assessing the spatial randomness

E. Dimensionality reduction techniques (PCA, t-SNE)

- **Explained Variance Ratio:** Measures the variance explained by each principal component in PCA.
  - This metric assesses the variance explained by each principal component, indicating how well the dimensionality reduction captures the variability in the Iris dataset.
- **Kullback-Leibler Divergence:** Evaluates the similarity between the high and low-dimensional data representation in t-SNE.
  - It evaluates the similarity between the high-dimensional data and the low-dimensional representation produced by t-SNE, providing insights into the quality of the dimensionality reduction.
- **Reconstruction Error:** Calculates the difference between the original and reconstructed data to evaluate the quality of dimensionality reduction.
- **Neighborhood Preservation:** Measures preserving pairwise distances or neighborhood relationships between data points in the high-dimensional and reduced-dimensional spaces.
- **Information Retention:** Besides explained variance, measures like Mutual Information or Normalized Mutual Information of retained information.

F. Outlier detection algorithms (Isolation Forest, LOF)

- **Precision and Recall:** Measure the accuracy of identifying outliers.
  - Precision measures the fraction of correctly identified outliers among the total predicted outliers, while recall calculates the fraction of correctly identified outliers among all actual outliers.
- **Receiver Operating Characteristic (ROC) Curve:** Plots the true positive rate against the false positive rate to evaluate the trade-off between sensitivity and specificity.
  - Illustrates the trade-off between the true positive rate and the false positive rate at various threshold settings, allowing for the evaluation of outlier detection performance.
- **Outlier Ranking:** Ranks the outliers based on their scores or anomaly scores assigned by the algorithm, allowing for prioritization or threshold selection.
- **Stability Analysis:** Assess the stability of outlier detection results by applying the algorithm on subsamples or different partitions of the data and comparing the consistency of the identified outliers.
- **Domain Expert Validation:** Collaborate with domain experts or use external sources of information to validate the detected outliers and ensure they align with domain knowledge.

ASSESS MODEL FIT ⇒ SUPERVISED

Evaluation methods encompass various aspects specific to each algorithm category. For clustering algorithms, the focus is on quantifying the formed clusters' quality, consistency, and discriminative ability. In dimensionality reduction techniques, the evaluation aims to assess information retention, preservation of data relationships, and visualization capabilities. For outlier detection, the evaluation methods gauge the accuracy and efficacy of identifying anomalies in the dataset.

SUPERVISED - Evaluation pitfalls

Mitigate pitfalls by employing a comprehensive evaluation, including cross-validation techniques, comparison to baselines or benchmarks, and consulting with domain expertise.

- **Overfitting -** supervised algorithms are prone to overfitting due to their ability to learn from the provided labeled data. Work to ensure algorithms measure unseen patterns beyond the evaluation dataset is essential.
- **Metric limitations** - Each evaluation metric has its limitations and underlying assumptions. Understand each measured characteristic and select those responding to the features of the dataset and project analysis objectives. Supervised algorithms require specific metrics tailored to the nature of the problem, such as accuracy, precision, recall, F1 score, ROC-AUC, or mean squared error (MSE). Consult each algorithm's research to understand its evaluation goals and characteristics.
- **Bias** - refers to situations where a dataset doesn't represent an underlying population's characteristics. To help address dataset bias in supervised algorithms, it is essential to consider issues related to biased sampling, class imbalance, or feature selection. Domain knowledge and thorough scientific research should guide identifying and mitigating dataset bias, including techniques like stratified sampling, resampling methods, or feature engineering.
- **Interpretation subjectivity** - understanding the contextual factors, domain specifics, and subjectivity involved in interpreting evaluation results is crucial. Collaboration with domain experts and considering the evaluation outcomes' practical implications can help make informed decisions.

SUPERVISED - Common Evaluation Metrics

I. **Accuracy:** Measures the overall correctness of the classification model's predictions.
II. **Precision:** Calculates the proportion of true positive predictions among the total positive predictions, indicating the model's ability to avoid false positives.
III. **Recall:** Computes the proportion of true positive predictions among the actual positive instances, measuring the model's ability to identify all relevant samples.
IV. **F1 Score:** Harmonic mean of precision and recall, providing a balanced measure of the model's performance.
V. **Area Under the ROC Curve (AUC-ROC):** Evaluate the model's ability to distinguish between different classes by plotting the true positive rate against the false positive rate.
VI. **Confusion Matrix:** Provides a tabular representation of the model's performance, showing the counts of true positive, true negative, false positive, and false negative predictions.

SUPERVISED - Classification Algorithms

Linear Regression

- Visualize Predictions: Plot the actual values against the predicted values to visually inspect the performance of the linear regression model.

Logistic Regression

- Receiver Operating Characteristic (ROC) Curve: Plot the ROC curve to evaluate the trade-off between true positive rate and false positive rate for different threshold settings.

Naïve Bayes

- No specific evaluation techniques listed.

k-Nearest Neighbors (k-NN)

- Receiver Operating Characteristic (ROC) Curve: Plot the ROC curve to evaluate the trade-off between true positive rate and false positive rate for different threshold settings.
- Hyperparameter Tuning: Iterate over different values for the 'k' parameter and evaluate the impact on the model's classification performance.

SUPERVISED - Consensus Algorithms:

Decision Trees

- Visualize the Tree: Plot the decision tree structure to gain insights into feature importance, decision paths, and interpretability.
- Feature Importance: Assess the importance of different features in the decision tree model using metrics such as Gini importance or information gain.

Random Forest

- Feature Importance: Determine feature importance based on the average impurity reduction or information gain across all decision trees in the random forest.
- Out-of-Bag Error: Utilize the out-of-bag error estimate as an additional evaluation metric to assess the model's performance without the need for separate validation data.

SUPERVISED - Linear Boundary Algorithms:

Support Vector Machines (SVM)

- Decision Boundary Visualization: Plot the decision boundary of the SVM model to visualize the separation between different classes.
- Hyperparameter Tuning: Iterate over different values for the hyperparameters (e.g., C, kernel type) and evaluate the impact on the model's classification performance.

Perceptrons

- No specific evaluation techniques listed.

SUPERVISED - Assessment Approaches

Residual Analysis:

- Perform a visual analysis of the residuals (the differences between the predicted and actual values) to assess the model's ability to capture the underlying patterns in the data. This can include plots such as residual vs. predicted values, residual vs. feature values, or a histogram of the residuals.

Cross-Validation

- Utilize cross-validation techniques, such as k-fold cross-validation, to estimate the model's performance on unseen data. This helps assess the generalization ability of the model and can provide more reliable performance estimates.

Learning Curves

- Plot learning curves to examine the model's performance as a function of the training data size. This can help identify issues such as underfitting or overfitting, as well as determine the sufficiency of the available training data.

Regularization Analysis

- If applicable to the supervised algorithm, analyze the effect of different regularization techniques (e.g., L1 regularization, L2 regularization) on the model's performance. This can involve tuning regularization parameters and evaluating the impact on the model's generalization ability.

Model Comparison

- Compare the performance of multiple supervised algorithms on the same dataset using appropriate evaluation metrics. This allows for a comparative analysis to identify the algorithm that best suits the specific problem and dataset.