# Module 5

## Evaluating Machine Learning Models

Supervised learning evaluation establishes how well a machine learning model can predict the outcome for unseen data.

It is essential for understanding model effectiveness and involves comparing model predictions to ground truth labels.

During training, the model tries to optimize predictions based on one or more evaluation metrics.

After training, the model is again evaluated to estimate how well it can generalize to unseen data.

Supervised learning evaluation is essential in both the training and testing phases.

Evaluating the performance of a machine learning model is **crucial** to understanding its ability to **generalize** to new data.

## 📌 Train-Test Split

The Train-Test Split technique is used to estimate how well a machine learning model will perform on unseen data.

📌 Why Split the Data?

✔ **Prevents Overfitting →** Ensures the model doesn’t memorize the training data.

✔ **Simulates Real-World Predictions →** Tests the model’s ability to generalize.

✔ **Avoids Data Snooping →** Prevents the model from learning patterns it shouldn't have access to.

📌 How It Works?

1. **Training Set →** Typically 70-80% of the dataset is used to train the model.
2. **Testing Set →** The remaining 20-30% is reserved for evaluation.

💡 Best Practice:

* Ensure the dataset is randomly shuffled before splitting.
* Use stratified sampling when dealing with imbalanced classes.

## 📌 Classification Metrics and Evaluation Techniques

A classification model predicts **categorical labels**. To assess how well the model performs, we use key **evaluation metrics**:

1. **Accuracy:**

**Definition:** The ratio of correctly predicted observations to the total number of observations

**Formula:**

✅ **Useful when** → Classes are **balanced** (i.e., similar numbers of positive and negative cases).

❌ **Not ideal when** → There is **class imbalance** (e.g., rare disease detection, fraud detection).

1. **Confusion Matrix:**

**Definition: confusion matrix** is a table that **compares the predicted class labels to the actual labels**. It helps visualize where the model is making **correct and incorrect predictions**.

|  |  |  |
| --- | --- | --- |
| Predicted  Actual | Positive | Negative |
| Positive | True Positive (TP) | False Negative (FN) |
| Negative | False Positive (FP) | True Negative (TN) |

* **True Positives (TP)** → Correctly predicted **positive** instances.
* **True Negatives (TN)** → Correctly predicted **negative** instances.
* **False Positives (FP)** → Incorrectly predicted **positive** instances (Type I Error).
* **False Negatives (FN)** → Incorrectly predicted **negative** instances (Type II Error).

1. **Precision: The Accuracy of Positive Predictions**

**Definition:** Precision measures how many of the predicted **positive** instances are **actually positive**.

**Formula:**

✅ **Use Precision when** → **False Positives** are **costly** (e.g., spam detection, recommending irrelevant ads).

1. **Recall (Sensitivity): Identifying Actual Positives**

**Definition:** Recall measures how many of the **actual positive instances** were **correctly predicted**.

**Formula:**

✅ **Use Recall when** → **False Negatives** are **critical** (e.g., missing a cancer diagnosis, fraud detection).

1. **F1 Score: The balance between Precision and Recall**

**Definition**: The **F1 Score** is the harmonic mean of **Precision and Recall**, providing a single measure of model effectiveness.

**Formula:**

✅ **Use F1 Score when** → Both False Positives and False Negatives are equally important.

### 🔹 Choosing the right metrict

|  |  |  |
| --- | --- | --- |
| **Scenario** | **Best Metric** | **Reason** |
| **Balanced Classes** | Accuracy | Works well when all classes have similar proportions. |
| **Class Imbalance** | Precision or Recall | Accuracy may be misleading if one class dominates. |
| **Medical Diagnosis** | Recall | False Negatives are more dangerous than False Positives. |
| **Spam Detection** | Precision | False Positives should be minimized. |
| **General Classification** | F1 Score | Balances Precision & Recall when both are important. |

## 📌 Regression Metrics and Evaluation Techniques

Evaluation is critical because:

* It quantifies how well the model generalizes to **unseen data**.
* It helps identify **underfitting, overfitting**, or potential **bias** in predictions.
* It guides **model selection, hyperparameter tuning**, and **interpretability**.

When a model makes predictions, the difference between the predicted and actual values is known as **error**. A well-performing regression model should minimize these errors.

### 🔹 Common Regression Evaluation Metrics

Regression metrics quantify **how far off** a model’s predictions are from the actual values. Each metric gives **different insights** into error magnitude, consistency, and model performance.

1. **MAE (Mean Absolute Error):**
   * Measures the average absolute difference between predicted values and actual values.
   * Every error is treated equally, regardless of direction or size.

**🔍 When to Use:**

* When **outliers are expected** but should not dominate the error measure.
* When **interpretability in actual units** (e.g., dollars, degrees) is important.

1. **Mean Squared Error (MSE):**
   * Squares the errors before averaging, thus **penalizing large deviations more heavily**.

**🔍 When to Use:**

* When **larger errors are unacceptable** and need to be discouraged during training.
* Ideal for algorithms that **optimize squared loss**, such as linear regression.

1. **Root Mean Squared Error (RMSE):**
   * The square root of MSE.
   * Expressed in the same units as the target variable, making it more intuitive than MSE.

**🔍 When to Use:**

* When you want an **error measure in the same unit** as your prediction target.
* Useful when large errors are very costly and should be emphasized.

1. **R² Score (Coefficient of Determination):**

* Measures the proportion of variance in the target variable that the model explains.
* Ranges from 0 (no explanatory power) to 1 (perfect prediction). Negative values indicates that model is **worse than just predicting the mean**.

**🔍 When to Use:**

* To assess the goodness of fit.
* To explain to stakeholders how much of the outcome your model captures.

ℹ️ Note: R² assumes **linear relationships** between features and target. In **nonlinear models**, it can be misleading if used in isolation.

## 📌 Evaluation insights

### 🔹 Combine Multiple Metrics for a Full Picture

* **MAE** gives an average error magnitude but is forgiving of outliers.
* **MSE/RMSE** emphasize larger errors, useful when **large mistakes are expensive**.
* **R²** tells how well the model captures the data structure but doesn’t reveal **error direction** or **bias**.

✔ Always **use multiple metrics together**, as no single number captures the whole performance landscape.

### 🔹 Use Residual Analysis to Detect Bias

A model might have strong metrics but still fail in specific ranges of the target variable. This is where residual plots become essential:

* Random scatter around zero → well-behaved model.
* Curved or sloped patterns → the model misses a nonlinear relationship.
* Systematic over/under-prediction in high or low ranges → distributional bias.

✔ Residual plots help identify where and why the model struggles, allowing targeted improvements.

### 🔹 High R² Can Be Deceptive

* A high R² (e.g., 0.85) doesn’t mean the model is **accurate everywhere**.
* It could just mean that the model predicts the **mean trend well**, but fails on the **extremes** or **outliers**.

✔ Visualizations (actual vs predicted, residuals) are crucial for verifying that high R² reflects **true performance**.

### 🔹 Systematic Error Across Ranges

Regression models often exhibit **non-uniform accuracy**:

* **Underprediction of high values** may indicate inability to extrapolate.
* **Overprediction of low values** might occur when the model is too biased toward the dataset mean.

✔ Address this by:

* Engineering features that **capture edge-case behavior**.
* Transforming the target variable (e.g., log, Box-Cox).
* Choosing a **model with more flexibility**, such as ensembles or gradient boosting.

### 🔹 Interpreting Feature Importance

When using models like **Random Forests**:

* **Feature importance** indicates **how much a feature reduces prediction error**, not its **causal effect**.
* Correlated features may **share importance**, which can dilute interpretability.

✔ Use a **correlation matrix** to confirm whether features are redundant or overlapping in information.

✔ Feature selection improves **model efficiency and robustness**.

### 🔹 Outliers and Clipping Impact Metrics

* Models trained on datasets with **target caps or skew** (e.g., price capped at $500,000) often underperform on **high-end predictions**.
* These effects are visible in residuals and degrade R² and RMSE selectively.

✔ Recognize when the **target distribution itself is biased or censored**, and adjust model goals or training data accordingly.

## 📌 Evaluating Unsupervised Learning, Heuristics and Techniques.

Evaluating unsupervised learning models is a different challenge compared to supervised ones. In supervised learning, we have labeled data to compare our predictions against — we know the correct answers. But in unsupervised learning, like clustering or dimensionality reduction, there’s no ground truth. These models try to **find patterns or groupings on their own**, which means we need different strategies to assess whether the patterns found are actually meaningful.

The purpose of unsupervised models is to **detect useful structure in data** — for example, identifying clusters or reducing dimensions in a way that keeps the data informative. Evaluation helps us answer questions like:

* Do the clusters represent real patterns in the data?
* Are similar data points grouped together consistently?
* When we reduce the number of dimensions, are we still keeping the important relationships between points?

Since we don’t have labels to directly measure against, we need to use tools like **heuristics**, **internal and external metrics**, **visualizations**, or **domain knowledge** to assess the quality of the results.

**🧠 Stability and Generalization – Key Concepts**

A key idea in unsupervised model evaluation is **stability** and **generalization**.

✔ **Stability:** A model produces similar results even when the dataset is slightly changed. For example, if you remove a few rows or shuffle the data, and the clusters still look mostly the same, that means the model is stable and reliable.

✔ **Generalization:** How well the model's discovered patterns hold up on new or different data. A good clustering model, for example, should form similar groupings even if the data varies a little. This is crucial for trusting the model’s insights, especially when we’re applying it to real-world, messy data.

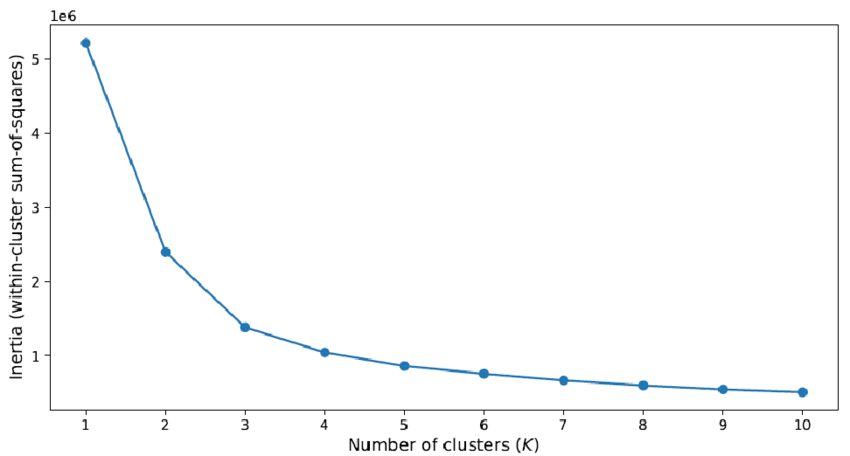
### 🔹 Internal Evaluation Metrics

These metrics use only the data and the clustering structure (no ground truth labels).

* 1. **Inertia:**

Measures the **sum of squared distances** between data points and cluster centers (used in k-means).

✔ Lower inertia indicates **tighter clusters**, but decreases as **k increases** — it needs to be interpreted with caution.

✔ Score can be plotted for different k values, and from it we can see how this score varies when k (number of clusters) change. Can be used to obtain the optimal number of clusters.

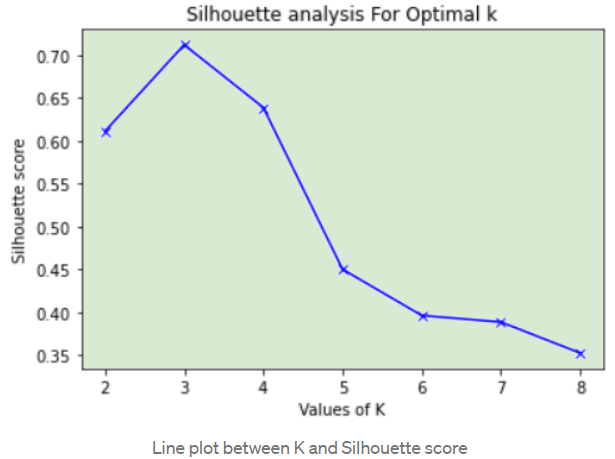
* 1. **Silhouette Score:**

Measures how close a point is compared to the rest of points that belongs to the same cluster, in relation to how close the same point is to the point in the nearest cluster.

It compares the two distance measures to see how similar a point is to its own cluster, compared to others.

✔ Score ranges from **-1 to 1** (higher values mean better-defines clusters).

✔ Great for **visualizing** and **validating** cluster **compactness** and **separation.**

✔ Score can be plotted for different k values, and from it we can see how this score varies when k (number of clusters) change. Can be used to obtain the optimal number of clusters.

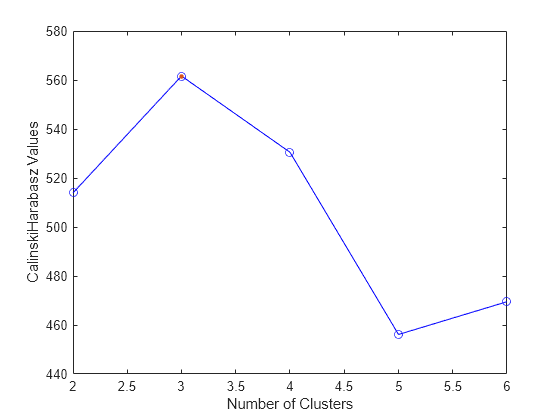
* 1. **Calinski-Harabasz Index (Variance Ratio Criterion):**

Measures the relationship, or ratio, of between-cluster dispersion (external distance) to within-cluster dispersion (internal distance).

✔ A well-defined cluster structure has a **high between-cluster dispersion** and **low within-cluster dispersion**.

✔ Scales well and is **computationally efficient**, making it ideal for larger datasets.

⚠️ Can sometimes **favor models with more clusters**, so best used in combination with other metrics.

✔ Score can be plotted for different k values, and from it we can see how this score varies when k (number of clusters) change. Can be used to obtain the optimal number of clusters.

* 1. **Davies-Bouldin Index:**

Measures **intra-cluster similarity** relative to **inter-cluster separation**.

✔ Lower values are better (les overlap between clusters).

* 1. **Dunn Index:**

Ratio of the **smallest distance between cluster centers** to the **largest intra-cluster distance.**

✔ **Higher values** mean better-defined, more isolated clusters.

✔ Sensitive to noise but provides strong insight into compactness and separation.

### 🔹 External Evaluation Metrics

External evaluation metrics are used when **true labels are available**, even though the learning process itself is unsupervised. These metrics help determine **how well the discovered clusters align with the actual class labels**.

They **compare the predicted cluster assignments** with the known categories (e.g., customer segments, document topics, species types), and provide **quantitative insight** into clustering quality based on **ground truth**.

These metrics are **especially valuable when validating a clustering algorithm** during benchmarking or research, or in semi-supervised learning scenarios.

* 1. **Adjusted Rand Index (ARI):**

Measures the **similarity between two assignments**: one from the clustering algorithm, and one from the actual labels.

✔ It **corrects for chance**, so random label assignments tend toward zero, while perfect alignment scores **1**.

✔ Negative values can occur if the clustering is **worse than random**.

How it works:

It looks at **pairs of samples** and checks how consistently they were placed together or separately in both the predicted and actual groupings.

Use it when:

✔ You want to test **how well your clustering matches known categories**.

✔ Evaluating robustness against **label shuffling** or **random noise**.

* 1. **Normalized Mutual Information (NMI):**

Measures the **amount of shared information** between the cluster assignments and true labels.

✔ The score ranges from **0 to 1**:

* 1 = perfect match (maximum information shared).
* **0** = completely unrelated assignments.

How it works:

Based on **entropy and information theory**, it quantifies how much knowing the predicted cluster helps in guessing the true class.

Normalization ensures fair comparison even when the number of clusters and classes differ.

Use it when:

✔ You want a **balanced, symmetric metric** that isn’t biased by the number of clusters or label distribution.

✔ Comparing different clustering configurations with uneven cluster sizes.

1. **Fowlkes-Mallows Index (FMI):**

Evaluates clustering by computing the **geometric mean of precision and recall** (based on pairwise point assignments).

✔ Ranges from **0 (no match)** to **1 (perfect clustering)**.

How it works:

Compares every pair of samples to see:

* Whether they belong to the **same cluster** in both true labels and predicted clusters.
* Whether they are correctly **separated or grouped**.

Use it when:

✔ It is sensitive to both false positives and false negatives in clustering.

✔ Provides a balanced view of clustering performance when both precision and recall are important.

### 🔹 Dimensionality Reduction Evaluation

Dimensionality reduction is often used to **project high-dimensional data into lower dimensions** — typically for **visualization, simplification, or preprocessing before clustering**. However, reducing dimensions always carries the risk of **information loss**, so it's crucial to **evaluate how much of the original structure is preserved** in the reduced space.

Evaluating a dimensionality reduction algorithm helps answer:

* Has the algorithm retained the **most important relationships between data points**?
* Can the reduced representation still support **meaningful clustering or classification**?
* Has it preserved **global and/or local structure**?

There are three key aspects to evaluate dimensionality reduction performance:

1. **Explained Variance Ratio (for PCA)**

* PCA transforms the data into **principal components**, which are ranked by how much **variance** (or spread) they explain from the original data.
* The **explained variance ratio** tells you how much information is retained in each principal component.
* If the **first few components** capture most of the variance, it means the reduced data still represents the original structure well.

✅ **Useful to decide how many dimensions you can safely reduce to without sacrificing too much information.**

🔎 *Example insight*: If the first 2 components explain 90% of the variance, you can confidently reduce your data to 2D for visualization or further analysis.

1. **Reconstruction Error**

* Applies to algorithms that **encode and decode data**, like PCA and autoencoders.
* Measures how well you can **reconstruct the original data** from the reduced representation.
* A **low reconstruction error** indicates that the reduced data still contains the essential information.

✅ **Used to assess how much information was lost** during the dimensionality reduction.

🔎 *Important when using reduced data for downstream tasks (e.g., predictions or clustering).*

1. **Neighborhood Preservation**

* Especially relevant for **nonlinear techniques** like t-SNE and UMAP.
* Measures how well the **local relationships between data points** (i.e., which points are close to which) are preserved in the low-dimensional space.
* This is crucial for visualizations or when clustering follows dimensionality reduction.

✅ **Evaluates whether nearby points in high-dimensional space remain nearby after reduction.**

🔎 *Important for tasks like manifold learning, anomaly detection, and cluster visualization.*