UNIVERSIDADE CATÓLICA PORTUGUESA

BRAGA

Machine Learning

Session 8 - T

Introduction to Supervised Learning

Ciência de Dados Aplicada

2023/2024

Unsupervised vs Supervised Learning

UNIVERSIDADE CATÓLICA PORTUGUESA
BRAGA

- **Unsupervised:** involves working with **unlabeled data**, where the algorithm explores the inherent **structure and patterns** within the input without explicit output guidance.

- **Supervised:** the algorithm is trained on a **labeled dataset**, where the input data is paired with corresponding output labels. The goal is to learn a **mapping from inputs to outputs**, allowing the algorithm to make predictions on new, unseen data.

Supervised Learning

- Given a **dataset**: $D = \{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$
  - where x_i represents input features and y_i represents corresponding labels.

- The goal is to learn a function $f(x)$ that **maps inputs to outputs**, i.e., $y_i = f(x_i) + \epsilon_i$.
  - Where ϵ_i represents a error term.

- While **minimizing the error** between the predicted output and real values.

Datasets for Supervised Learning

UNIVERSIDADE CATÓLICA PORTUGUESA
BRAGA

Inputs: **Output:**

Features →

Examples →

(...)

Mileage	Engine	Horsepower	Transmition Type	Car Type
25000	2.0	180	Manual	Sedan
30000	2.5	200	Automatic	SUV
20000	1.8	160	Manual	Sedan
35000	3.0	250	Automatic	SUV
28000	2.2	190	Automatic	Sedan
32000	2.8	220	Manual	SUV
27000	2.0	170	Manual	Sedan
...	...	...	...	...

Feature types → Continuous | Continuous | Continuous | Discrete | Discrete

Binary Classification

UNIVERSIDADE CATÓLICA PORTUGUESA
BRAGA

Inputs: Output:

Features → Examples → (...)

Mileage	Engine	Horsepower	Transmition Type	Car Type
25000	2.0	180	Manual	Sedan
30000	2.5	200	Automatic	SUV
20000	1.8	160	Manual	Sedan
35000	3.0	250	Automatic	SUV
28000	2.2	190	Automatic	Sedan
32000	2.8	220	Manual	SUV
27000	2.0	170	Manual	Sedan
...	...	...	...	...

Feature types → Continuous, Continuous, Continuous, Discrete, Discrete

→ **Binary Classification (0 or 1)**

Binary Classification

UNIVERSIDADE CATÓLICA PORTUGUESA
BRAGA

Inputs: **Output:**

Features →

Examples →

(...)

Mileage	Engine	Horsepower	Transmition Type	Brand
25000	2.0	180	Manual	Toyota
30000	2.5	200	Automatic	Ford
20000	1.8	160	Manual	Honda
35000	3.0	250	Automatic	BMW
28000	2.2	190	Automatic	Toyota
32000	2.8	220	Manual	Honda
27000	2.0	170	Manual	Ford
...	...	...	...	...

Feature types → Continuous | Continuous | Continuous | Discrete | Discrete

→ **Multiclass Classification (> 2 classes)**

Binary Classification

UNIVERSIDADE CATÓLICA PORTUGUESA
BRAGA

Inputs: **Output:**

Features →

Examples →

(...)

Mileage	Engine	Horsepower	Transmition Type	Price
25000	2.0	180	Manual	20000
30000	2.5	200	Automatic	25000
20000	1.8	160	Manual	18000
35000	3.0	250	Automatic	30000
28000	2.2	190	Automatic	28000
32000	2.8	220	Manual	15000
27000	2.0	170	Manual	13000
...	...	...	...	...

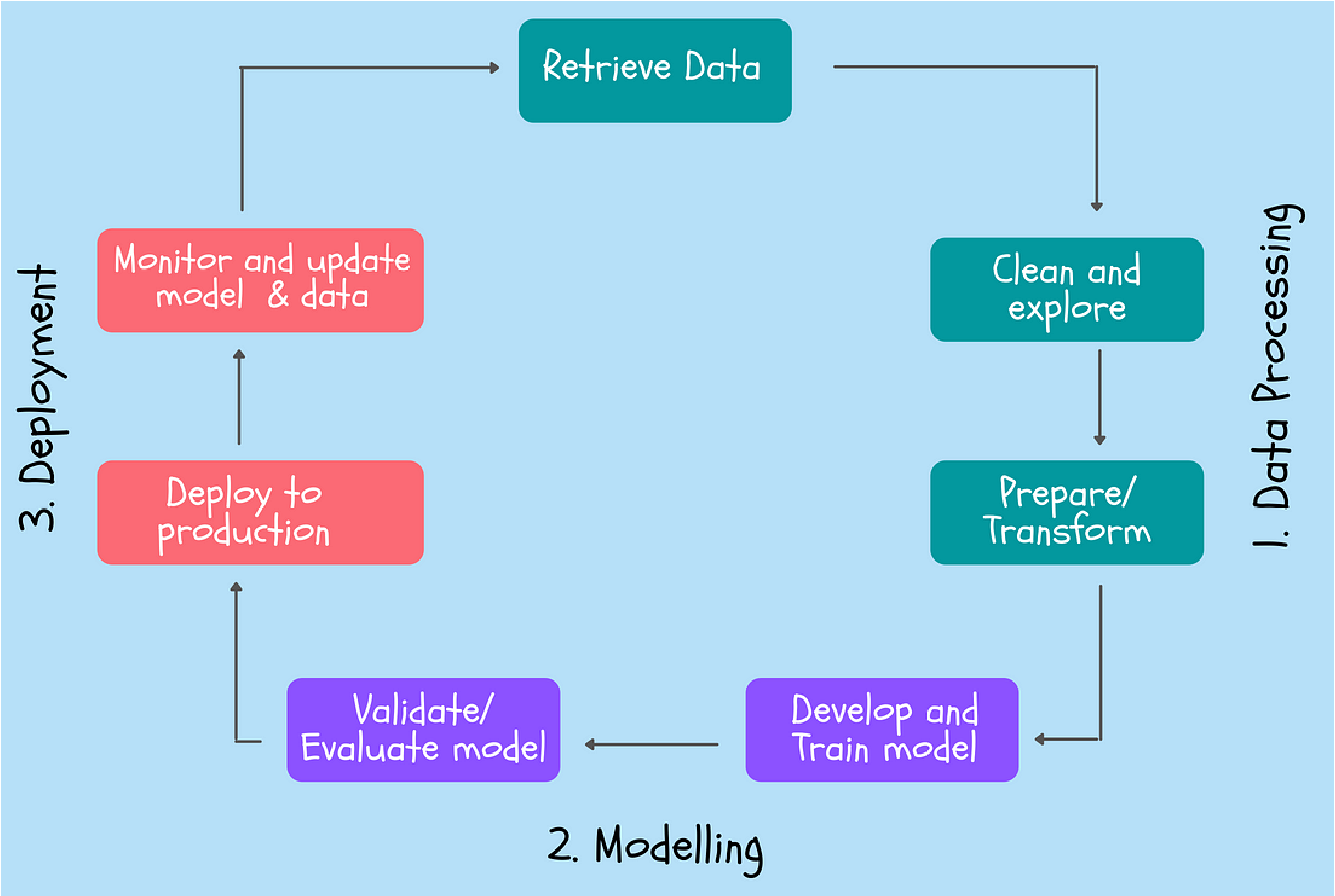
→ **Regression**

Feature types → Continuous Continuous Continuous Discrete Continuous

Supervised Learning or Not?

1. Predict the IMDB score of a movie based on its characteristics.
2. Identify the illness of a patient based on its symptoms.
3. Group patients based on the values of indicators from their biochemical analyses.
4. Predict the weather for October 2023 based on the weather of previous months.
5. Calculate the average age of the students in this course.
6. Writing a program to improve its performance when playing chess against humans.

Supervised Learning Workflow



https://towardsdatascience.com/the-machine-learning-workflow-explained-557abf882079

Supervised Learning Workflow

UNIVERSIDADE CATÓLICA PORTUGUESA
BRAGA

- **Prepare the data:**
  - Data collection;
  - Data cleaning;
  - Data preprocessing.

- **Model building:**
  - Selecting the model;
  - Model architecture;
  - Choose hyperparameters.

- **Train and evaluate the model:**
  - Train the model with the training data to minimize a loss function;
  - Assess the model's performance on a separate validation set to tune hyperparameters and prevent overfitting.
  - Assess the model's performance on the test.

- **Get predictions from the model:**
  - Use the trained model to make predictions on new, unseen data.

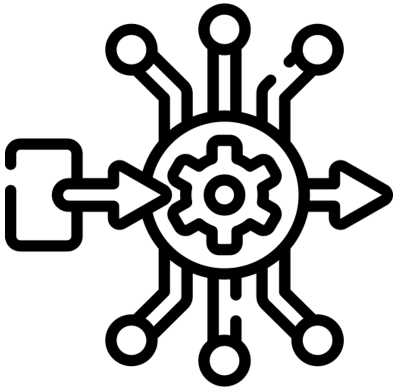
Supervised Learning Workflow

Dataset

Mileage	Engine	Horsepower	Transmition Type	Car Type
25000	2.0	180	Manual	Sedan
30000	2.5	200	Automatic	SUV
20000	1.8	160	Manual	Sedan
35000	3.0	250	Automatic	SUV
28000	2.2	190	Automatic	Sedan
32000	2.8	220	Manual	SUV
27000	2.0	170	Manual	Sedan
...	...	...	...	...

Optimization Algorithm

Model



Linear Models
Tree-Based Models
Instance-Based Models
Probabilistic Models
Ensemble Models
Neural Networks
...

Mileage	Engine	Horsepower	Transmition Type	Car Type
25000	2.0	180	Manual	?
30000	2.5	200	Automatic	?
20000	1.8	160	Manual	?
...	...	...	...	...

Predictions

Car Type
SUV
SUV
Sedan
...

Model Evaluation: Error Metrics

- Assessing the quality of a model for a specific task involves the computation of **error metrics**.

- These metrics provide insights into **how well the model performs** on a set of examples (not used during model training).

- The metric to use depends on the **type of problem**: regression or classification.

Classification Metrics

- **Confusion matrix:**

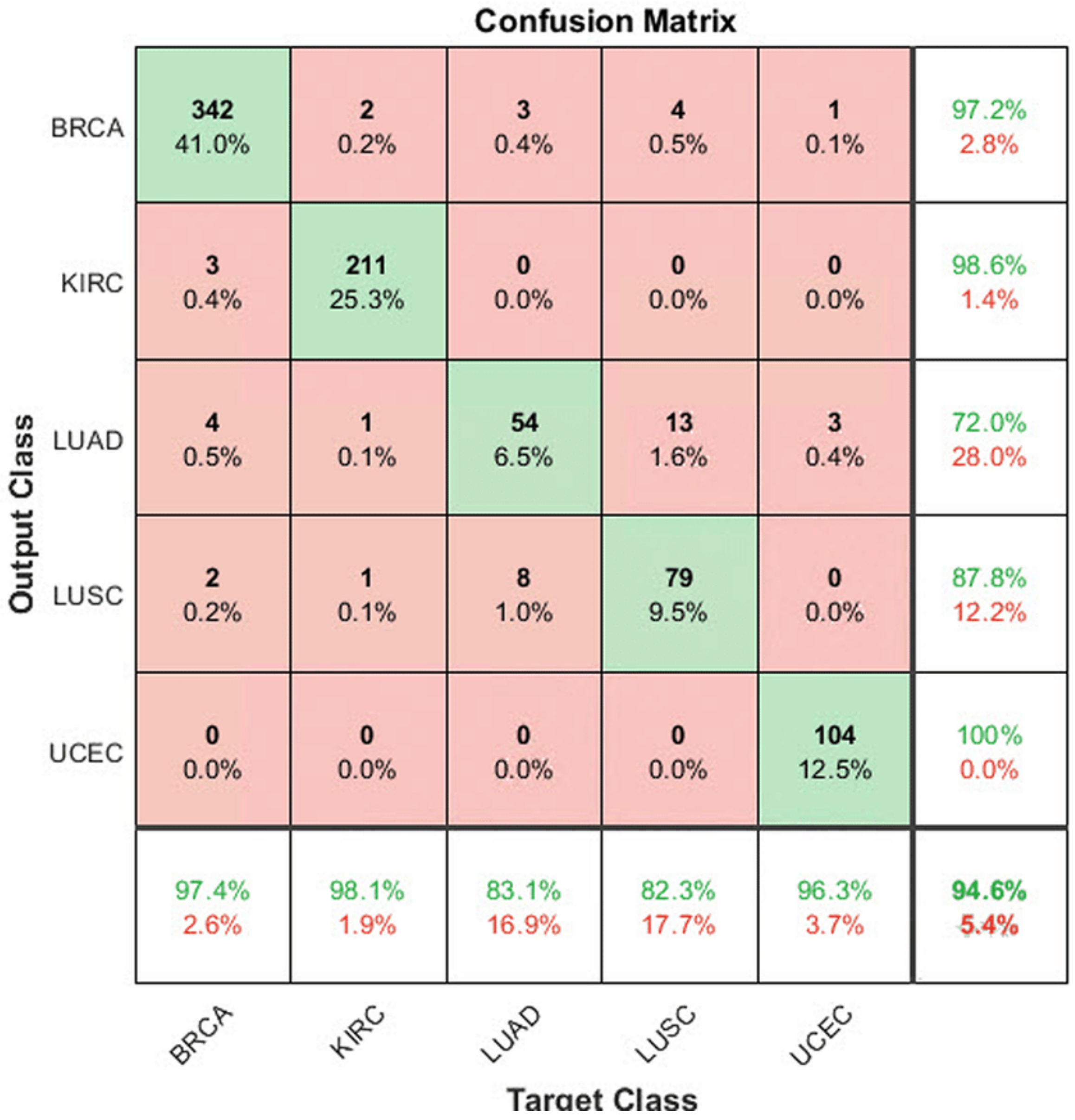
2 classes

		Predicted Class	
		Positive	**Negative**
Actual Class	**Positive**	True Positive (TP)	False Negative (FN) **Type II Error**
	Negative	False Positive (FP) **Type I Error**	True Negative (TN)

Classification Metrics

- **Confusion matrix:**

More than 2 classes



Classification Metrics

aka **Recall**

		Predicted Class		
		Positive	**Negative**	
Actual Class	**Positive**	True Positive (TP)	False Negative (FN) **Type II Error**	**Sensitivity** $\frac{TP}{(TP+FN)}$
	Negative	False Positive (FP) **Type I Error**	True Negative (TN)	**Specificity** $\frac{TN}{(TN+FP)}$
		Precision $\frac{TP}{(TP+FP)}$	**Negative Predictive Value** $\frac{TN}{(TN+FN)}$	**Accuracy** $\frac{TP+TN}{(TP+TN+FP+FN)}$

Classification Metrics

$$\text{F1 Score} = \frac{2}{\frac{1}{\text{Precision}} + \frac{1}{\text{Recall}}} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Matthews Correlation Coefficient $$\text{MCC} = \frac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{\sqrt{(\text{TP}+\text{FP})(\text{TP}+\text{FN})(\text{TN}+\text{FP})(\text{TN}+\text{FN})}}$$

Classification Metrics

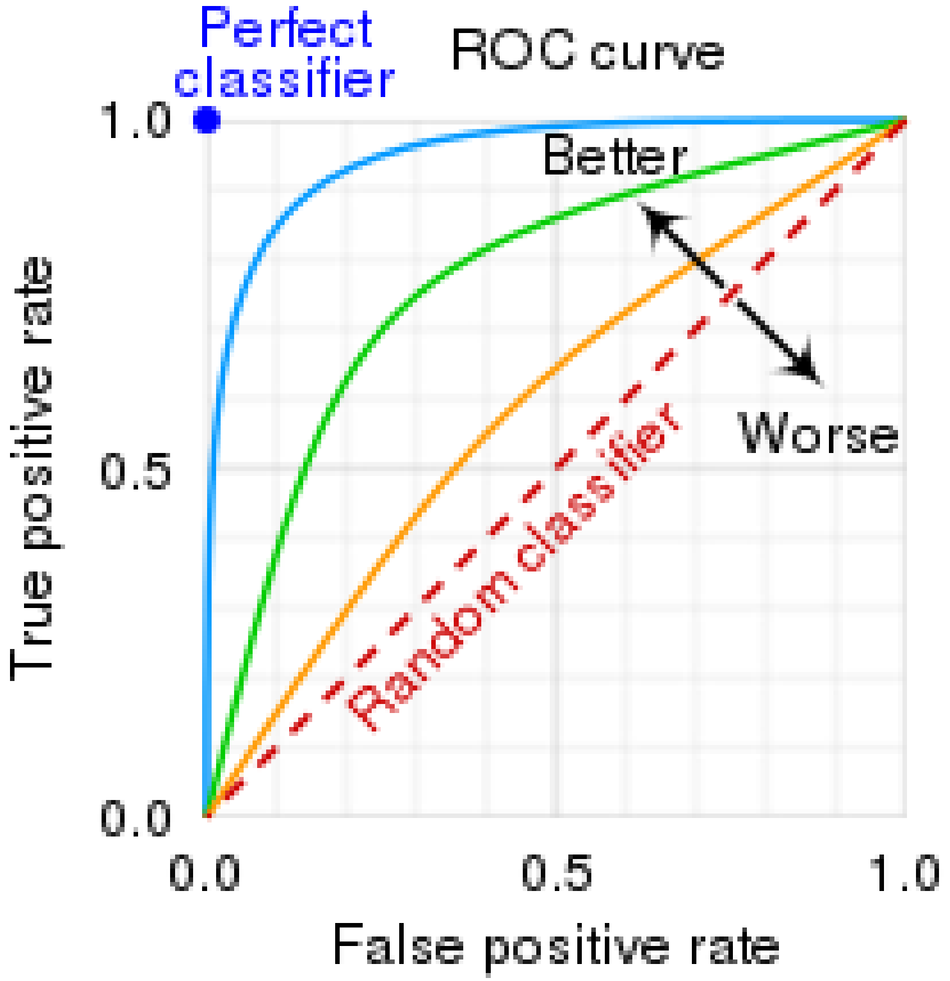
- **Receiver operating characteristic (ROC) curves**

- Graphically evaluates model discrimination across **different thresholds**.

- True Positive Rate vs. False Positive Rate

- Area Under the Curve (**AUC**): Reliable classifier quality indicator (0.5 for random, 1 for perfect).

- **Precision-Recall curves**: More suitable for **imbalanced data**, highlighting precision-recall trade-offs.



Classification Metrics: Which one to pick?

UNIVERSIDADE CATÓLICA PORTUGUESA
BRAGA

- **Precision:** general measure used when the **classes are balanced** and misclassifications of both positive and negative cases are equally important;
- **Sensitivity/Recall:** when correctly identifying **positive cases** is crucial (e.g. medical diagnosis or fraud detection);
- **Specificity:** when correctly identifying **negative cases** is important (e.g. security screening or quality control);
- **Precision:** when we want to **minimize false positives** (e.g. email spam fetection);
- **F1-score:** when we want a **balance between precision and recall**, especially in situations where there is an **imbalance** between the number of positive and negative cases.
- **MCC:** when you want a single metric that considers the **overall performance** of the model, especially in **binary classification** tasks where the classes are **imbalanced**.

Regression Metrics

UNIVERSIDADE CATÓLICA PORTUGUESA
BRAGA

- **Mean Absolute Error** (MAE):
  $$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}|$$
  - Simple and interpretable measure of average prediction error. Less sensitive to outliers.
- **Mean Squared Error** (MSE):
  $$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y})^2$$
  - Penalize larger errors more heavily. Sensitive to outliers due to squaring the errors.
- **Root Mean Squared Error** (RMSE):
  $$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y})^2}$$
  - Metric in the same units as the target variable. Provides an interpretable measure like MAE but accounts for larger errors like MSE.

Regression Metrics

UNIVERSIDADE CATÓLICA PORTUGUESA
BRAGA

- **Mean Absolute Percentage Error** (MAPE): $\mathrm{MAPE} = \frac{1}{\mathrm{n}} \sum_{i=1}^{n} \frac{|\mathrm{y_i} - \hat{\mathrm{y}}|}{\mathrm{y_i}} * 100\%$
  - Useful when the scale of the target variable varies widely.

- **Coefficient of Determination** (R-squared): $R^2 = 1 - \frac{\sum(y_i - \hat{y})^2}{\sum(y_i - \bar{y})^2}$
  - Used to assess how well independent variables explain variability in the dependent variable. Higher values indicate a better fit of the model to the data.

- **Adjusted R-squared**: $R^2_{adj} = 1 - \left[\frac{(1 - R^2)(n-1)}{n-k-1}\right]$
  - Adjusts the R-squared for the number of independent variables (k), providing a more accurate reflection of model fit.
  - n is the number of observations in the data.

Error Estimation Methods

UNIVERSIDADE CATÓLICA PORTUGUESA
BRAGA

- **Objective:** ensure **credible evaluation** of algorithm performance and **generalization ability**.

- Error measures should not be applied to the same dataset that was used for training.

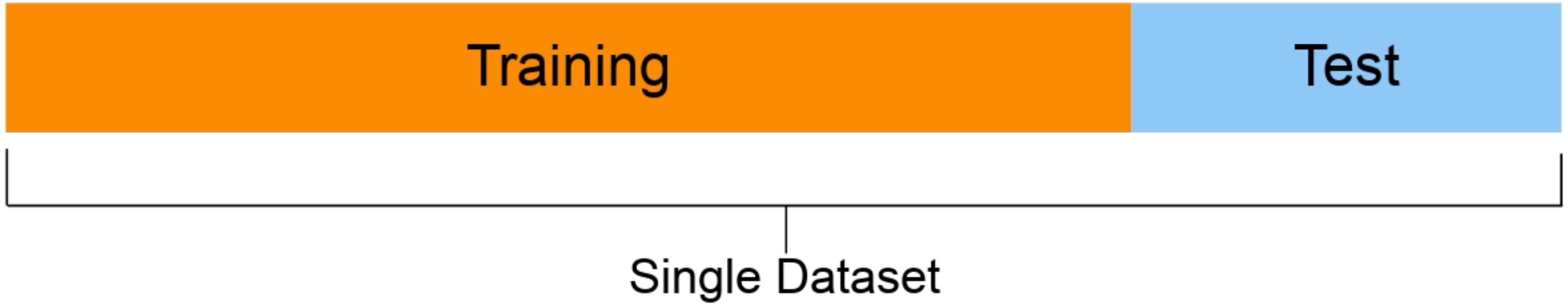
- **Validation and test sets** are used to evaluate the trained model.

- Importance of Test Examples:
  - Crucial for assessing how well the model generalizes to **unseen data**.
  - Ensures **unbiased evaluation** of model performance.

Holdout

- It involves splitting the dataset into two subsets: the **training set** and the **test set**.

- The model is trained on the training set and evaluated on the independent test set.



Holdout

- Sometimes, it is necessary to split the data into three subsets: the **training set**, the **validation set**, and the **test set**.
- The model is trained on the training set and evaluated on the validation set to **tune hyperparameters**.
- Finally, the model's performance is assessed on the test set to obtain an unbiased estimate of its **generalization ability**.

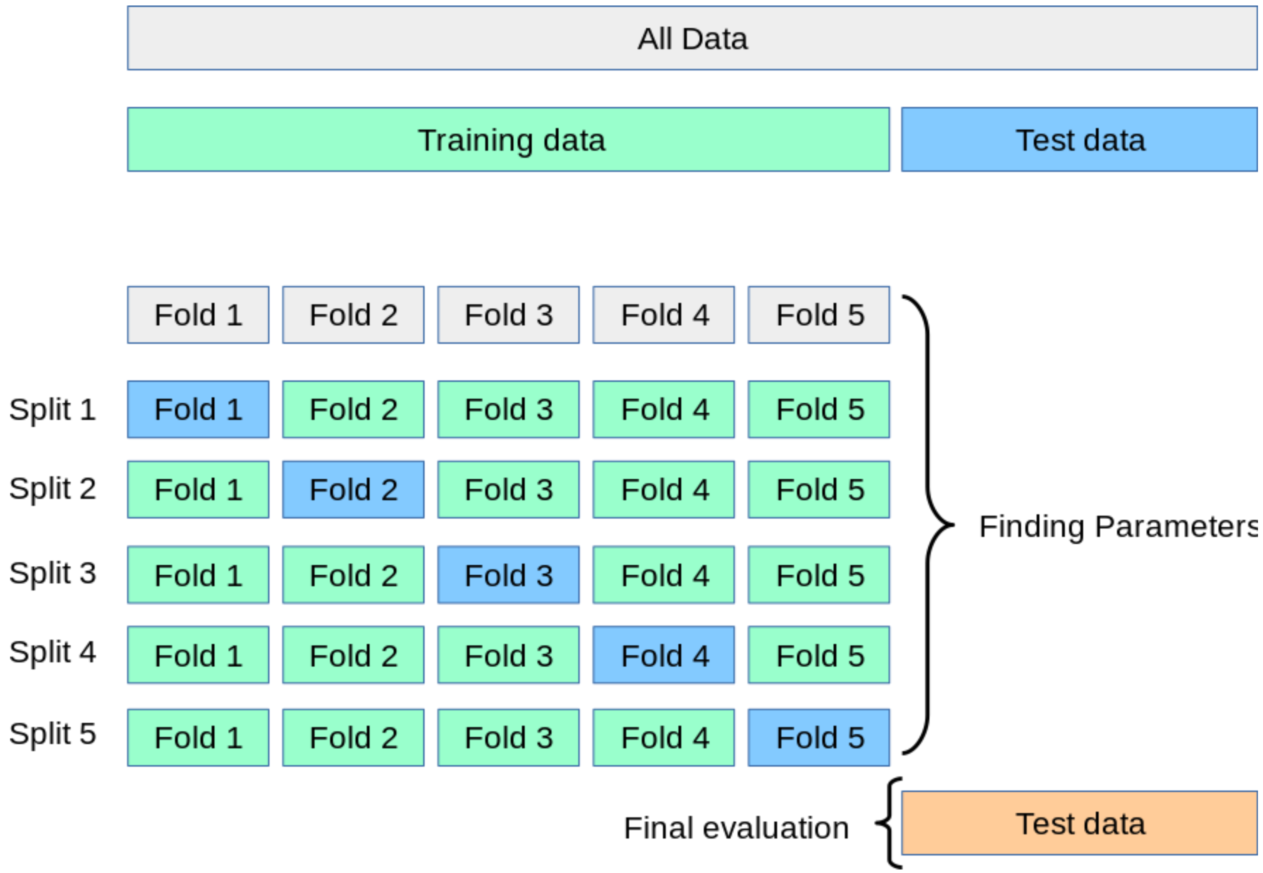
Holdout

- **Advantages:**
  - Easy to implement.
  - Provides a quick estimate of model performance.
  - Useful for large datasets where computational resources are limited.

- **Limitations:**
  - Performance estimate may vary depending on the random split of data.
  - May not be suitable for small datasets due to potential data imbalance.

Cross Validation

- Cross-validation is a robust technique for estimating prediction error by iteratively splitting the dataset into multiple subsets.



Cross Validation

UNIVERSIDADE CATÓLICA PORTUGUESA
BRAGA

- Types of Cross-Validation:
  - **K-Fold Cross-Validation:** Divides the data into k folds, each used as a test set once.
  - **Leave-One-Out Cross-Validation (LOOCV):** Each observation is used as a test set once, with the rest as the training set (k=number of samples).

- **Advantages:**
  - Provides a robust estimate of model performance by averaging over multiple iterations.

- **Limitations:**
  - Computationally intensive, especially for large datasets or complex models.
  - May result in higher variance estimates due to randomness in data splits.

Learning Bias

UNIVERSIDADE CATÓLICA PORTUGUESA
BRAGA

- It represents the **systematic error or deviation** of the model's predictions from the true values.

- Learning bias can arise due to model complexity, insufficient data, or inherent limitations of the algorithm.

- Types of learning bias:
  - **Underfitting (High Bias)**
  - **Overfitting (Low Bias, High Variance)**

Bias and Variance

- **Bias:**
  - Bias refers to the error introduced by approximating a real-world problem with a simplified model.
  - High bias models are too simple and fail to capture the underlying patterns in the data.

- **Variance:**
  - Variance measures the model's sensitivity to small fluctuations in the training data.
  - High variance models are overly complex and capture noise or random fluctuations in the data.

- **Bias-Variance Tradeoff:**
  - Find a tradeoff between bias and variance: reducing one typically increases the other.
  - The goal is to find the right balance that minimizes both bias and variance, resulting in optimal model performance.

Underfitting

- Underfitting occurs when a machine learning model **fails to capture the underlying patterns in the data**, resulting in poor performance on both training and test data.

- **Causes:**
  - Model complexity is too low relative to the complexity of the underlying data.
  - Insufficient features or training examples to capture the variability in the data.

- **Mitigation Strategies:**
  - Increase model complexity by adding more features or using a more complex algorithm.
  - Fine-tune hyperparameters to achieve a better balance between bias and variance.
  - Collect more data to provide the model with a richer learning environment.

Overfitting

UNIVERSIDADE CATÓLICA PORTUGUESA
BRAGA

- Overfitting occurs when a machine learning model **captures noise or irrelevant patterns from the training data**, leading to poor generalization on unseen data.

- **Causes:**
  - Model complexity is too high relative to the amount of training data available.
  - Too many features or interactions being considered, leading to capturing noise instead of signal.

- **Mitigation Strategies:**
  - Simplify the model by reducing complexity, such as decreasing the number of features or using regularization techniques.
  - Increase training data to provide the model with more diverse examples.
  - Use techniques like cross-validation to evaluate model performance and select the best-performing model.

Resources

- Kelleher, J. D., Namee, B. M., & D’Arcy, A. (2015). Fundamentals of machine learning for predictive data analytics. London, England: MIT Press.

- https://courses.washington.edu/me333afe/Bias_Variance_Tradeoff.pdf