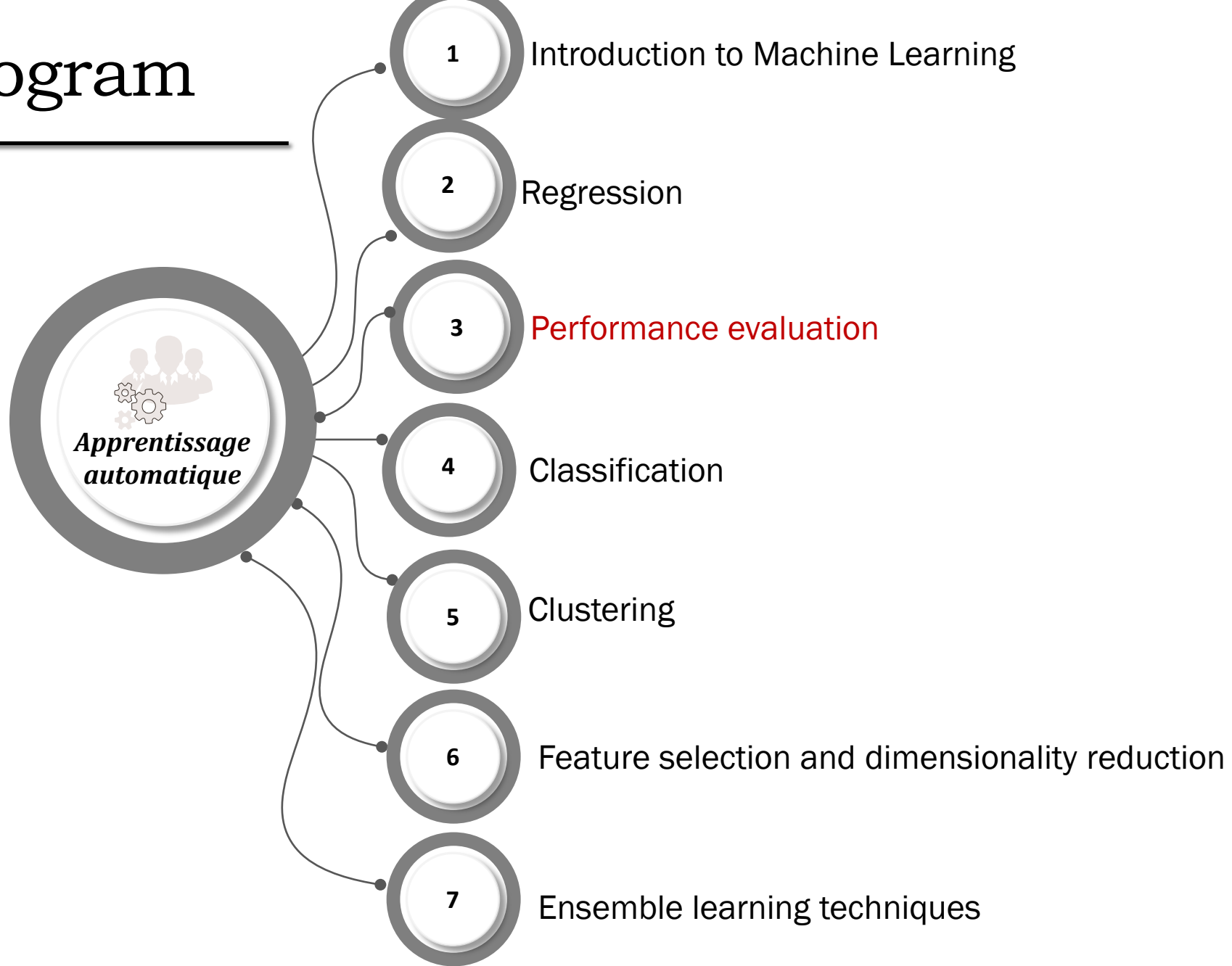


Machine Learning

*Intelligence Artificielle et
Sciences de Données
(IASD)*

DR N. DIF

Program



1. EVALUATION METRICS

- The data mining process is composed of three main steps: data preprocessing, modeling, and performance evaluation.
- Generally, **the performance** of the generated models is computed to compare between the generated models from training.
- The evaluation step is crucial in the development of predictive models in machine learning because the **main purpose of training is not only to create predictive models but rather to create high-performing models** characterized by promising results in prediction.





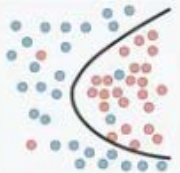
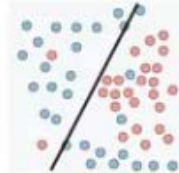

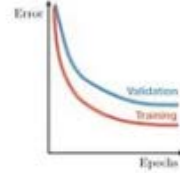

2. EVALUATION METRICS

- Evaluation metrics are used to measure the trained model's performance on the test set.
- There are two types of evaluation metrics: to maximize such as accuracy and to minimize such as error. The choice of the appropriate metric depends on the application domain and the dataset's nature.
- $\text{Accuracy} = \frac{\text{Number of correctly classified samples}}{\text{Total number of samples}}$, $\text{Error} = \frac{\text{Number of misclassified instances}}{\text{Total number of samples}}$



How many evaluation measures are there in regression?

2. OVERFITTING AND UNDERFITTING

	Overfitting	Generalization	Underfitting
Regression Illustration			
Classification Illustration			
Deep Learning Illustration			

Overfitting : The model performs well on the training data but poorly generalize on the test data. This results in a significant gap between the performance on the training and test datasets.

Underfitting : The model performs poorly on both the training and test datasets.

From : <https://medium.com/analytics-vidhya/how-to-evaluate-your-machine-learning-model-76a7671e9f2e>

2. OVERFITTING AND UNDERFITTING

- A good quality training dataset should respect three keys : quality, quantity, and variability.
 1. **Quality** : the considered samples should not be wrong and should have a good quality (treat missing and outliers), In short, if you do not provide the right data to your ML models, you will get wrong results.
 2. **Quantity** : a good dataset is characterized by a large number of samples, where a maximum of cases is considered
 3. **Variability** : In real-world problems, it's impossible to consider all possible scenarios. The solution is to ensure generalization by varying in the presented samples.
- A high-performing model is not only characterized by good performance but by a strong generalization on unseen data during training.
- Overfitting is the opposite of generalization that defines models characterized by a high variance between performance on the training data and unseen samples during training.

3. EVALUATION TECHNIQUES



How to split the dataset?

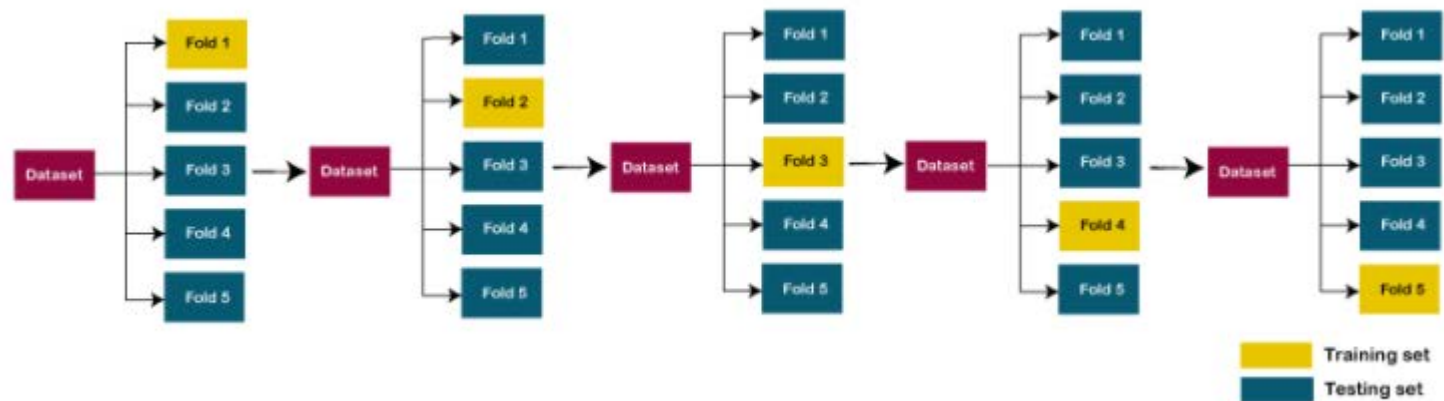
- It is important to measure the model's performance on unseen data during training to avoid overfitting problems and to ensure generalization.
- Evaluation methods are among the used techniques used to prevent overfitting. Generally, the dataset is split into a training and a test sets.
- The training set is used to train the model, while the test set is used to assess its performance. In some situations, the dataset is split into three sets: a training set, a validation set, and a test set. The validation set is used during training to fine-tune the values of some hyperparameters.

3. EVALUATION TECHNIQUES

- **Hold out:** in this technique, the dataset is randomly divided into a training set and a test set (or a validation set), typically $2/3$ of the dataset is used for training and the remaining $1/3$ for testing. This approach is not recommended for unbalanced datasets or datasets characterized by attributes associated with predominant values.
- **Stratified Hold out:** it addresses the issue of unbalanced datasets by equally dividing the classes or feature values in the training and test sets. The drawback of this approach is the number of instances considered during training and testing, as an algorithm that consider a large number of samples during training generates a model characterized by good generalization. On the other hand, to assess the performance of the generated model, it is important to test it on a large number of samples.

3. EVALUATION TECHNIQUES

- **K-Cross-Validation** : This evaluation method addresses the problems of stratified hold-out. It suggests to split the dataset into K folds, then k-1 folds are used for training and 1 fold for testing (or evaluation). The process is repeated K times, and then the performance is computed based on the average test performances on each fold. Generally, $k = 10$



3. EVALUATION TECHNIQUES

- **Stratified K-cross-validation:** this technique is similar to K-fold cross-validation, except that it exploits the principle of fair splitting to ensure that each group represents of the entire dataset (same principle as stratified hold out). It is one of the best used approaches for to handle low bias and high variance problems.
- **Leave-one-out :** it uses the same principle as the K-fold cross-validation with $K = N-1$, where N is the number of samples in the dataset. This technique uses only one sample for testing in each iteration, and the number of iterations present the number of samples N in the dataset.