MACHINE LEARNING

Classification et Clustering

# Référents module : Jérémie Suzan et Théo Trouillon

# 

# Objectifs

A l’issue de ce module, vous serez capable de :

* Entrainer et évaluer un modèle de classification
* Utiliser des méthodes d’ensemble
* Mettre en évidence les phénomènes de sur/sous apprentissage
* Entrainer et évaluer un modèle de clustering

# Pré-requis

* Programmation en Python
* Bases de statistique
* Régression linéaire

# Projet étape 1 : Classification (1 jour)

### Modalités

* Travail en autonomie
* Production individuelle

### Compétences

* Se familiariser avec la bibliothèque scikit-learn
* Savoir entraîner un modèle de classification et faire des prédictions
* Connaître les différentes métriques d’évaluation pour les problèmes de classification
* Mettre en place une procédure de sélection de modèle par grid-search et cross-validation

### Consignes

* Ouvrir et compléter le notebook

### Ressources

* <https://scikit-learn.org/stable/tutorial/basic/tutorial.html>
* <https://scikit-learn.org/stable/auto_examples/classification/plot_classifier_comparison.html>
* <https://scikit-learn.org/stable/modules/cross_validation.html>
* *“Hands on machine learning …”*, chapitres 2 et 3
* *“Introduction to statistical learning”*, chapitre

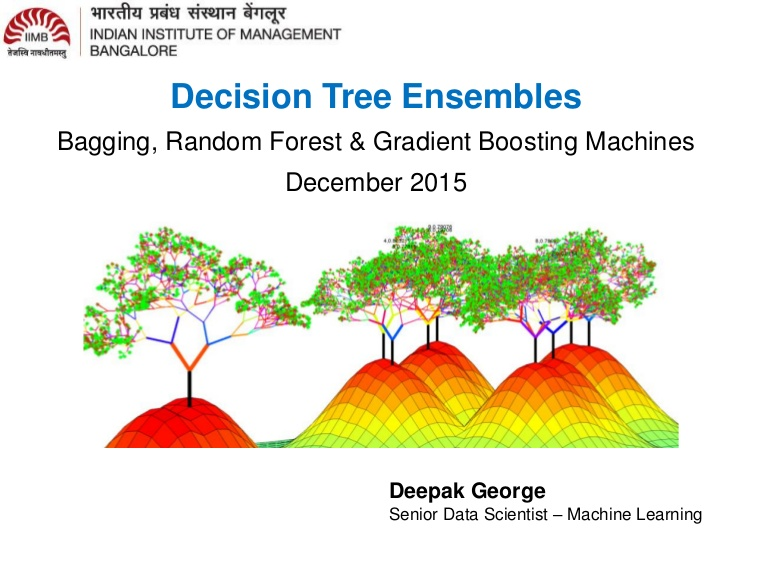
### Livrables

* Répondre aux questions du fichier mémo
* Le notebook rempli, permettant d’évaluer les performances d’un classifieur par k plus proches voisins

# 

# 

# Projet étape 2 : Introduction aux méthodes d’ensemble (1,2 jour)



### Modalités

* Travail en autonomie
* Production individuelle

### Compétences

* Entraîner un modèle de classification en utilisant les techniques de bagging et de boosting.
* Trier les paramètres d'un problèmes par ordre d'importance.
* Évaluer les performances d'un modèle de classification.

### Consignes

* Téléchargez l’archive contenant le projet (un notebook et un jeu de données)
* Compléter le notebook

### Ressources

* <https://scikit-learn.org/stable/modules/ensemble.html>
* <https://martin-thoma.com/ensembles/>
* <https://medium.com/@rrfd/boosting-bagging-and-stacking-ensemble-methods-with-sklearn-and-mlens-a455c0c982de>
* <https://xgboost.readthedocs.io/en/latest/index.html>
* <https://www.lpsm.paris/pageperso/has/source/Hand-on-ML.pdf> *(chapitre 7)*

### Livrables

* Visualisation du classement des paramètres sous forme d’histogramme
* Utilisation des méthodes d’ensemble
  + Notebook complété.
  + Mémo/Schéma sur les méthodes d’ensemble comprenant:
    - Schéma de fonctionnement des méthodes de bagging et de boosting.
    - Avantage/Inconvénients de chacunes des méthodes.

# Projet étape 3 : Introduction au partitionnement (clustering) (0,8 jour)

### Modalités

* Travail en autonomie
* Production individuelle

### Compétences

* Utiliser des méthodes de partitionnement
* Trouver le nombre de cluster optimal
* Créer des partitions à partir d’un jeu de données en utilisant des méthodes mise à disposition dans scikit-learn
* Visualiser les partitions créées par un algorithme de partitionnement

### Consignes

* Compléter le notebook

### Ressources

* [https://scikit-learn.org/stable/modules/clustering.html#clustering](https://scikit-learn.org/stable/modules/clustering.html" \l "clustering)
* <https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html>
* <https://scikit-image.org/>

### Livrables

* Analyses de silhouette commentée de résultats de clusterings qui doit comprendre:
  + Une silhouette avec des clusters non uniformes.
  + Une silhouette avec des clusters uniformes.