

# Introduction au Machine Learning

## Organisation et critère d'évaluation des mini-projets

### *English version in p.3 and 4*

#### 1. Organisation générale

L'évaluation du cours « Data Analytics & Machine Learning » est effectuée pour 70% de la note finale par un mini-projet obligatoire.

Ce mini-projet utilisera les bases de données disponibles et téléchargeables sur le site : <https://archive.ics.uci.edu/datasets>. Pour chacune, la base de données est accompagnée d'une explication des variables et des types de problèmes d'intérêt, ainsi qu'une liste d'articles scientifiques souvent eux aussi téléchargeables qui utilisent la base en question.

Le travail demandé consiste en les activités suivantes :

- Constitution des groupes de 3 à 5 étudiants – il n'y a pas de restriction sur la composition des groupes ;
- Choix d'une base de données qui servira d'exemple d'application pour le mini-projet (**différente pour chaque groupe !!**) ;
- Utilisation de la base de données pour résoudre un problème pertinent : apprentissage, classification, régression, clustering... :
  - o Formulation du problème ;
  - o Codage de l'algorithme de résolution choisi (l'utilisation de Matlab ou de Python est autorisée au choix des étudiants) ;
  - o Tests et validation des solutions, comparaison des résultats en fonction des réglages choisis ;
  - o Réflexions sur les avantages, limites et perspectives sur l'approche proposée.
- Lecture et synthèse d'un des articles illustrant par ailleurs l'utilisation de la base de données ;
- Rédaction d'un rapport (15 pages maximum) incluant le travail réalisé sur la base de données et la présentation de l'article choisi. Ce rapport devra inclure une page « Mode d'emploi » permettant d'utiliser facilement les codes fournis et de les tester pour différents réglages.

## 2. Deadline

- La déclaration de la constitution des groupes et le choix de la base de données doit être fait avant le **lundi 13 octobre 2025 à 12h**. Cette déclaration doit être faite en remplissant le document :

[https://docs.google.com/spreadsheets/d/1E14s1FsG7pjFOAraQ3a28W2glZAGafTISYtt\\_fS5n4c/edit?gid=1773951741#gid=1773951741](https://docs.google.com/spreadsheets/d/1E14s1FsG7pjFOAraQ3a28W2glZAGafTISYtt_fS5n4c/edit?gid=1773951741#gid=1773951741)

- **Il convient de choisir des bases de données différentes pour chaque groupe.**
- Le rendu des livrables devra être effectué par mail à [guillaume.sandou@centralesupelec.fr](mailto:guillaume.sandou@centralesupelec.fr) sous la forme d'une archive unique incluant les programmes et le rapport en respectant la nomenclature DAML\_GroupeN. La deadline de remise est le **jeudi 13 novembre 2025 à 23h59**, avec un délai de courtoisie exceptionnel au **dimanche 16 novembre à 23h59**.

*(Pour rappel, la période filière débute le lundi 17 novembre 2025, période filière durant laquelle vous n'êtes pas censés travailler pour les activités de la mention – il n'y aura donc aucune prolongation supplémentaire de la deadline).*

## 3. Attendus et barème

Le module Data Analytics & Machine Learning représente un volume horaire de 21 heures en présentiel. Selon le référentiel en vigueur dans l'école, cela doit correspondre à un travail personnel d'environ 15h à 20h incluant le travail sur le cours entre les séances, les révisions pour l'examen QCM du 05/11 et ce mini-projet. Le mini-projet doit donc correspondre à un travail d'environ 8 à 10h **par étudiant**.

- Travail technique – **12 pts**
  - Pertinence du problème étudié – **1pt**
  - Complexité de la base de données choisie – **2 pts**
  - Formulation mathématique du problème – **2pts**
  - Codage commenté de la solution au problème – **3 pts**
  - Clarté et pertinence des visualisations choisies pour la présentation des résultats – **1pt**
  - Comparaison et analyse des résultats pour différents réglages / variantes des algorithmes – **2 pts**
  - Analyse des avantages, limites et perspectives de l'approche proposée – **1pt**
- Rapport (15 pages maximum – inutile de mettre des codes en annexe) – **8 pts**
  - Clarté de la rédaction et des illustrations – **3pts**
  - Synthèse de l'article scientifique (compréhension, analyse et mise en perspective avec le travail effectué) – **4pts**
  - Mode d'emploi utilisateur – **1 pt**

# Introduction to Machine Learning

## Organization and evaluation criteria for mini-projects

### 1. Organisation générale

The evaluation of the “Data Analytics & Machine Learning” course is carried out for 70% of the final grade by a compulsory mini-project.

This mini-project will use the databases available and downloadable on the site: <https://archive.ics.uci.edu/datasets>. For each, the database is accompanied by an explanation of the variables and types of problems of interest, as well as a list of scientific articles, often also downloadable, which use the database in question.

The work required consists of the following activities:

- Constitution of groups of 3 to 5 students – there is no restriction on the composition of groups;
- Choice of a database which will serve as an example of application for the mini-project (**different for each group!!**);
- Use of the database to solve a relevant problem: learning, classification, regression, clustering, etc.:
  - Formulation of the problem;
  - Coding of the chosen resolution algorithm (the use of Matlab or Python is authorized at the students' choice);
  - Testing and validation of solutions, comparison of results according to the chosen settings;
  - Reflections on the advantages, limitations and perspectives on the proposed approach.
- Reading and synthesis of one of the articles illustrating the use of the database;
- Writing a report (15 pages maximum) including the work carried out on the database and the presentation of the chosen article. This report must include a “How to Use” page allowing you to easily use the codes provided and test them for different settings.

### 2. Deadline

- The declaration of the constitution of the groups and the choice of the database must be made before **Monday October 13, 2025 at noon**. This declaration must be made by completing the document:

[https://docs.google.com/spreadsheets/d/1El4s1FsG7pjFOAraQ3a28W2glZAGafTISYtt\\_fS5n4c/edit?gid=1773951741#gid=1773951741](https://docs.google.com/spreadsheets/d/1El4s1FsG7pjFOAraQ3a28W2glZAGafTISYtt_fS5n4c/edit?gid=1773951741#gid=1773951741)

- **Different databases should be chosen for each group.**
- The deliverables must be rendered by email to [guillaume.sandou@centralesupelec.fr](mailto:guillaume.sandou@centralesupelec.fr) in the form of a single archive including the programs and the report, respecting the DAML\_GroupeN nomenclature. The submission deadline is **Thursday November 13, 2025 at 11:59 p.m.**, with an exceptional courtesy deadline of **Sunday November 16 at 11:59 p.m.**

*(As a reminder, the “filières” period begins on Monday November 17, 2025, a period during which you are not supposed to work for the activities of the mention - there will therefore be no further extension of the deadline).*

### **3. Expectations and grading**

The Data Analytics & Machine Learning module represents an hourly on-site volume of 21 hours. According to the standard rules in the school, this must correspond to personal work of approximately 15 hours to 20 hours including work on the course between sessions, revisions for the MCQ exam of 05/11 and this mini-project. The mini-project must therefore correspond to approximately 8 to 10 hours of work **per student**.

- Technical work – **12 pts**
  - Relevance of the problem studied – **1pt**
  - Complexity of the chosen database – **2 pts**
  - Mathematical formulation of the problem – **2pts**
  - Commented coding of the solution to the problem – **3 pts**
  - Clarity and relevance of the visualizations chosen for the presentation of the results – **1pt**
  - Comparison and analysis of results for different settings / variants of algorithms – **2 pts**
  - Analysis of the advantages, limits and prospects of the proposed approach – **1pt**
- Report (15 pages maximum – no need to append codes) – **8 pts**
  - Clarity of writing and illustrations – **3pts**
  - Summary of the scientific article (understanding, analysis and perspective with the work carried out) – **4pts**
  - User instructions – **1 pt**