Nédra Mellouli 2025

GUIDELINES
MACHINE LEARNING PROJECT

Choice of the topic

Grade details - Code

Grade details – Report

Grade details – Video

Datasets examples

Project organization

CHOICE OF TOPIC

- **Identify a theme** that genuinely interests you and is related to your **field of specialization**. → Think about problems you are curious about or areas where you would like to gain more expertise.
- **Consult** your professors/associations: Discuss your initial ideas with your major/Lab supervisor, or with associations, or others. They can provide guidance, suggest directions, and help refine your research question.
- **Identify data sources**: Search for open and publicly available datasets that can be used to address your project question. Consider data quality, accessibility, and relevance to your chosen theme. Example of sources:
  - Kaggle
  - Google Dataset Search
  - UCI Machine Learning Repository
  - GitHub
  - Etc…

GRADE DETAILS - CODE

- Importation of libraries and dataset
- Dataset analysis : number of columns/rows, number of values, variables definition, quality, visualization, correlation analysis, data balance, etc…
- Data pre-processing : missing values, duplicates, inconsistencies, outliers, encoding, scaling, train-test-split…
- Creation of the first models + model’s hyperparameter’s tuning
- Address overfitting, underfitting or other obstacles in your project
- Relevant metrics
- Dimension reduction
- Ensemble models + more advanced models (use scientific papers as reference)
- Comparison of models and conclusion
- Overall coherence in the code + comments
- **List of bonus / malus :**
  - ALL your lines of codes and comments must be in **english**, otherwise
  - If your code is especially well structured.

GRADE DETAILS - REPORT

- Business case : ddefine precisely the problem you want to solve, the objective of your project, and how this is directly linked to your field of specialization. This will directly influence the type of data you need to collect.
- Description of your dataset and the source
- Data exploration, graphics and figures about your dataset
- Formalization of your problem
- Presentation of your models
- How you addressed each obstacle you encountered (parameter optimization, overfitting, underfitting, data unbalance, etc…)
- Comparison of models results
- Conclusion about how you tackled your business case
- Scientific papers, references, external sources
- **List of bonus / malus :**
  - ALL your report must be in **english**, otherwise your grade will be **DIVIDED BY 2**.
  - If your report is difficult to read or to understand
  - If your report is especially well written
  - If you wrote your report using LaTex, +1 pts or +2 pts according to the quality of your work.

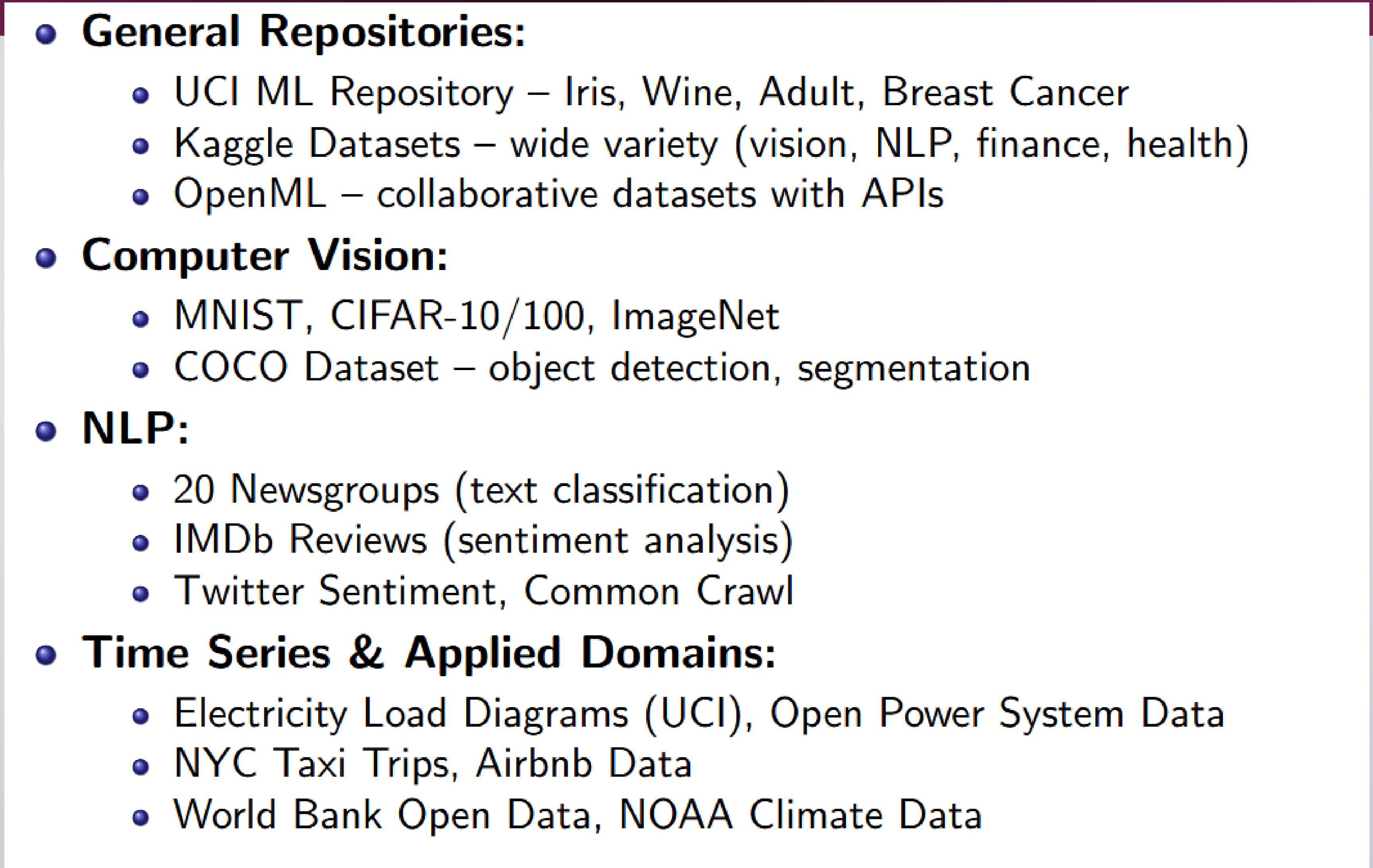
GRADE DETAILS - VIDEO

- Timing of the video (between 4 and 5 minutes)
- Presentation of the business case and dataset (~1 min)
- Presentation of your models (~2 min)
- Explanation of obstacles and project evolution (~1 min)
- Conclusion about the project and the results (~1 min)
- Overall clarity
- Good overview of the code
- **List of bonus / malus :**
  - The whole video must be in **english**, your grade will be **DIVIDED BY 2**.
  - If your level of English is especially good
  - If you made all the members of your group participate

EXAMPLES OF DATASETS

@ALL SPECIALISATIONS

EXAMPLE OF DATA SOURCES



- **General Repositories:**
  - UCI ML Repository – Iris, Wine, Adult, Breast Cancer
  - Kaggle Datasets – wide variety (vision, NLP, finance, health)
  - OpenML – collaborative datasets with APIs
- **Computer Vision:**
  - MNIST, CIFAR-10/100, ImageNet
  - COCO Dataset – object detection, segmentation
- **NLP:**
  - 20 Newsgroups (text classification)
  - IMDb Reviews (sentiment analysis)
  - Twitter Sentiment, Common Crawl
- **Time Series & Applied Domains:**
  - Electricity Load Diagrams (UCI), Open Power System Data
  - NYC Taxi Trips, Airbnb Data
  - World Bank Open Data, NOAA Climate Data

ENERGIE & INDUSTRIE

- **Electricity Load Diagrams (UCI)** – séries temporelles de consommation.
- **Open Power System Data** (https://open-power-system-data.org/
- **PJM Hourly Energy Consumption** (Kaggle).

TRANSPORT & IOT

- **NYC Taxi Trips** (https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page)
- OpenStreetMap – données géospatiales libres
- Airbnb Data (InsideAirbnb) – locations et prix.

ENVIRONNEMENT & CLIMATE

- **NOAA Climate Data** (https://www.ncei.noaa.gov/)
- **World Bank Open Data** (https://data.worldbank.org/)
- **Global Forest Watch** – données satellites sur la déforestation.

FINANCE & ECONOMY

- **Credit Card Fraud Detection** (Kaggle).
- **Lending Club Loan Data** (Kaggle).
- **Quandl** (https://www.quandl.com/)

HEALTHCARE & MEDECINE

- **Breast Cancer Wisconsin Dataset** (UCI) – tumor diagnosis.
- **COVID-19 Open Research Dataset (CORD-19)** .
- **MIMIC-III** (controlled Access) – data ICU

ROBOTIC

- **Robotic Vision (Perception, SLAM, Navigation)**
- **KITTI Vision Benchmark Suite**
  - Autonomous driving dataset with stereo images, LiDAR, odometry, and object detection.
  - http://www.cvlibs.net/datasets/kitti/
- **TUM RGB-D Dataset**
- RGB-D sequences from a Kinect camera for SLAM and visual odometry.
- https://vision.in.tum.de/data/datasets/rgbd-dataset
- **EuRoC MAV Dataset**
- Drone (MAV) indoor flights with synchronized cameras, IMU, and ground-truth trajectories.
- https://projects.asl.ethz.ch/datasets/doku.php?id=kmavvisualinertialdatasets

OCC/CCC

- **Malware Analysis**
- **EMBER (Endgame Malware Benchmark for Research)**
  - Windows PE malware dataset with extracted features for ML.
  - https://github.com/endgameinc/ember
- **Malicia Dataset**
- Collection of malicious binaries (mainly Windows malware).
- https://www.stratosphereips.org/datasets-malware
- **VirusShare / VirusTotal (API access)**
- Large repositories of malware samples.
- https://virusshare.com/

IOT SECURITY

- **TON_IoT Datasets**
  - IoT telemetry, operating system logs, and network traffic (benign + cyber attacks).
  - https://research.unsw.edu.au/projects/toniot-datasets
- **BoT-IoT Dataset**
- Large-scale dataset of botnet attacks in IoT networks.
- https://research.unsw.edu.au/projects/bot-iot-dataset

INTRUSION

- **Intrusion Detection / Network Traffic**
- **KDD Cup 1999 (KDD99)**
  - Classic intrusion detection dataset (old but still used for benchmarks).
  - Contains simulated attacks (DoS, U2R, R2L, probing).
  - http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html
- **NSL-KDD**
- Improved version of KDD99 (removes redundant records, more balanced).
- https://www.unb.ca/cic/datasets/nsl.html
- **CICIDS2017**
- Modern dataset with realistic traffic (Benign + attacks: DDoS, Brute Force, Botnet, Web Attacks).
- Created by Canadian Institute for Cybersecurity (CIC).
- https://www.unb.ca/cic/datasets/ids-2017.html

PROJECT ORGANIZATION: 4 STEPS

RECOMMANDATIONS FOR STUDENTS

Home Work by a team of 3 students

STEP 1: PRE-PROJECT

- Students must be in groups of no more than 3
- Students must choose a subject related to their major. To do this, they should exchange with the heads of their majors (with whom I'm already in discussion), their tutor, Kaggle and of course ChatGPT to learn and not to copy!
- Groups have two weeks (starting from the CMO on date of 09/24/2025) to define the business objectives and the scope.

STEP 2: IMPLEMENTATION OF STANDARD SOLUTIONS

1. Analyse the data (check quality, statistical information, define variables, imbalancing data, correlation analysis, reduction, etc) and pre-processing
2. Implement solutions for each task using the algos seen in class
3. Define the learning and testing plan: choice of data, construction of sets and methods for controlling overfitting
4. Critical analysis of the results using evaluation metrics

STEP 3: IMPROVING THE STANDARD SOLUTION

1. Implement advanced versions of the algorithms seen in class
2. Define the learning and testing plan: choice of data, construction of sets and methods for controlling overfitting
3. Analyze and critique your results using evaluation metrics
4. Combine several algorithms for ensemble learning decision making
5. Choose an algorithm outside the scope of the course, which may even be deep learning
6. Explanation of the algorithm and justification of the choice with a scientific paper that serves as a reference (example: articles on Google Scholar)
7. Implementation of the algorithm, evaluation and comparison with previous results and also with the baseline.

STRUCTURE OF THE REPORT

JUST AN EXAMPLE TO HELP YOU!

DOCUMENT STRUCTURE

- Business scope
- Problem formalisation and methods
  - Algorithm description
  - Limitations
- Methodology
  - Data description and exploration
    - Missing values
    - Imbalanced data
    - Outliers
  - Data splitting for train/test
  - Algorithm implementation and hyperpameters
- Results
  - Metrics
  - Overfitting/underfitting/Imbalance…
  - Evaluation, comparison with the baseline
- Discussion and conclusion

DEADLINES AND DELIVRABLES

- The Lab 8: Evaluation of the project progress -> Explanantion of your code in progress
- The Lab 9: Final Evaluation of the project with oral presentation of 10mn by group.
- 23 december: Final version (Github of the project) of code + Video