Rules for the project

Different steps

Step 1: Pre-project

- Students must be in groups of no more than 4
- 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!
- Groups have two weeks (starting from the CMO on date of 10/16/2024) to define the business objectives and the scope. This initial work gives rise to a 1st deliverable of two pages to be posted on moodle (I give the structure to follow).
- Date : first version 30/10 on DVL and last version 11/08.
- See the document structure on DVL

Stage 1: 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
5. Deliverable from stage 1: document deriving each task, its formalisation, methodology, results obtained and criticisms, conclusion and prospects

Date Stage 1:

Date: 7/11: for ACT2; FIN;MMN3;MMN4;OCC2

Date 14/11: IF4 ; IND;MMN2;

Date 15/11: CCC1;CCC2,DIA1; DIA 3; DIA4;DIA 6; EVD2; IF2,IF3;IF5, MMN1;SB

DIA5 20/11 3h

DIA2 22/11 3h

Date 22/11: EV1;EVD3;IF1;OCC1;OCC3 ;

Date 29/11: ACT1

Stage 2: 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

Deliverable for stage 2: document highlighting the limitations of the algorithms, the solutions chosen with an explanation of the algorithms, the results obtained and comparison with previous results, discussion and conclusion

Dates Stage 2:

Date 20/11 DIA5;MMN4

Date 21/11 CCC1; CCC2;

Date 22/11 DIA2;DIA4;DIA6;EVD2;MMN1;

Date 27/11 IND;

Date 28/11 IF4;MMN2;MMN3;OCC2;

Date 29/11 ACT2;DIA1;DIA3;EVD1;FIN;IF1;IF2;IF3;IF5;OCC1;SB

Date 4/12 EVD3

Date 5/12 ACT

Date 11/12 OCC3

Stage 3 : more improvements, “more and more”

1. Choose an algorithm outside the scope of the course, which may even be deep learning
2. Explanation of the algorithm and justification of the choice with a scientific paper that serves as a reference (example: articles on Google Scholar)
3. Implementation of the algorithm, evaluation and comparison with previous results and also with kaggle if the project is inspired by it.
4. Stage 3 deliverable (Final document): Description of the project, context and objectives, formalization of the problem, methodology, data and results, discussion and conclusion + References (bibliography).

Dates Stage 3:

- Date 4/12 IND
- Date 5/12 MMN2;MMN3;OCC2;
- Date 6/12 ACT2;DIA1;DIA2;DIA3;DIA4;DIA6;EVD1;EVD2;FIN;IF2;IF3;IF5;MMN1;OCC1;SB
- Date 11/12 ACT1;DIA4;EVD3;MMN4;OCC3; DIA5
- Date 12/12 CCC1;CCC2;IF4;

Evaluation

Evaluation

- Intermediate Evaluation of stage 1 work in-progress according to Date Stage 1

- Intermediate Evaluation of stage 2 work in-progress according to Date Stage 2

- Intermediate Evaluation of stage 3 work in-progress according to Date Stage 3

- Post the final document, sources and a video of your oral presentation (5mn) at 26/12/2024 in DVL

Machine Learning Grade

Pre-project(use case)	**Projet**	**PW(TP)**	**Quiz**	**total**
20%	40%	15%	25%	100
Projet	Stage 1	Stage 2	Stage 3	
	0.3	0.3	0.4	

- Pre-project: It replaces the use case. The evaluation of pre-project will take into account the following criterias

Criteria	Business challenge description	Data description & sources	Scope of the project	Work plan	Grade (total)
Points	30	30	30	10	
Description	-Identify, select and critically analyze different specialized resources to document a business subject aligned with the major -synthesize the resources and define the further use case perimeter	-Description of the data -Mention of data origin, data sources	-Understand and analyze the complexity of the project scope, -Formalize the problem -define the spectrum of possible models to design possible solutions	-Define a project timetable-Define the contribution of each member of the group	100 points

- stage 1: its evaluation will take into account the progress of the solutions and the following criterias are considered

Criteria	Problem formalization and ML tasks identification	Choice and motivation of algorithms	Algorithms Description and hyperparameters	Data description & shortcomings	Methodology	Results & evaluation	Code	Grade (total)
Points	5	15	5	25	25	15	10	
Description	-Able to understand the problem and express it with machine learning objectives	-Able to choose the right machine learning models to achieve these objectives -Able to criticise the limitations of the models selected, which can anticipate the quality of the expected results	-Explain algorithms -Explain the choice of the hyperparameters	Identify the data shortcomings and use appropriate techniques to highlight theses problems	-Data pre-processing -data splitting on train (validation)/test datasets -implementation of algorithms and code quality, -hyperparameters optimisation -monitoring of under-/over-fitting -early stopping	-define metrics to evaluate the performance of algorithms -critical analysis of results and areas for improvement	Comment your code!	100 points And the coef is 0.3

- stage 2: its evaluation will take into account the progress of the solutions and the following criterias are considered

Criteria	Improvement Assumptions	Choice and motivation of algorithms	Algorithms Description and hyperparameters	Methodology	Results & evaluation	code	Grade (total)
Points	5	15	5	30	25	20	
Description	-Able to understand the lack of data and models express new assumptions with machine learning objectives	-Able to choose the right machine learning models to achieve these objectives -Able to criticise the limitations of the models selected, which can anticipate the quality of the expected results	-Explain algorithms -Explain the choice of the hyperparameters -Explain the complexity of each algorithm	-Data pre-processing (if different from the previous stage) -data splitting on train (validation)/test datasets (if the strategy is different from the previous stage) -implementation of algorithms and code quality, -hyperparameters optimisation -monitoring of under-/over-fitting -early stopping	-define metrics to evaluate the performance of algorithms -critical analysis of results and areas for improvement -complexity evaluation and computing time	-comment your code	100 points And the coef is 0.3

- stage 3: its evaluation will take into account the progress of the solutions and the following criterias are considered

Criteria	Improvement Assumptions	Choice and motivation of algorithms	Algorithms Description and hyperparameters	Methodology	Results evaluation	Video	Code	Grade (total)
Points	5	10	5	15	20	30	15	
Description	-Able to understand the lack of data and models express new assumptions with machine learning objectives	-Able to choose the right machine learning models to achieve these objectives -Able to criticise the limitations of the models selected, which can anticipate the quality of the expected results	-Explain algorithms -Explain the choice of the hyperparameters -Explain	-Data pre-processing -data splitting on train (validation)/test datasets -implementation of algorithms and code quality, -hyperparameters optimisation -monitoring of under-/over-fitting -early stopping	-define metrics to evaluate the performance of algorithms -critical analysis of results and areas for improvement	-Respect the time, -pedagogical presentation	Comment your code!	100 points And the coef is 0.4

Example

Example: train circulation topic

- We collected data every five minutes, on train circulation: predicted arrival delay at each station (compared to official timetable).
- Data source : https://www.kaggle.com/datasets/bartek358/train-delays
- No more information available than what is explained on this link
- The aim of your pre-project step is to propose uses cases. For train circulation problem we can develop what can be achieve with such a data set ? We try to think as our customer do. So as an initial request, as your customer for this project, could be “to predict for each train it’s real arrival time”

Methodology(1):

- You have to prepare 3 sprint review
  - Share suggestion on what can be done with the dataset with your lecturer (or customer)
  - Prepare questions to transform your initial request in a real project.
  - The last sprint review you will fix your ideas. It may be not definitive but should give a real and concrete insight of where the project might land.

Documents structure

Document (2 pages) content

- The core of the document shoud have these three topics: Business ; Data ; Objectives or targets.
- Business: you need to think about real business topics or known (or unknown) challanges and how you can generate values by using machine learning process.
- Data: it's a question of identifying the data you need to work on
- Objectives: List the functionalities to be achieved and for each functionality you will associate a task.

1rst Document structure (pre-project)

- Business challange and state-of-the-art
- Data description and data sources
- Business objectives and the scope
- Work plan
- Conclusion
- References

2nd Document structure (Stage 1)

- 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
    - Evaluation
- Discussion and conclusion

3rd Document structure (Stage 2)

- Previous methods and limitations
- Improvement Assumptions
- Problem formalisation and methods
    - New Algorithm description
    - Limitations
- Methodology
    - New Algorithm implementation and hyperpameters
- Results
    - Metrics
    - Overfitting
    - Evaluation and comparision with previous solutions
- Discussion and conclusion

4rd Document structure (Stage 3)

- Previous methods and limitations
- Improvement Assumptions
- Problem formalisation and methods
    - New Algorithm description
    - Limitations
- Methodology
    - New Algorithm implementation and hyperpameters
- Results
    - Metrics
    - Overfitting
    - Evaluation and comparision with previous solutions
- Discussion and conclusion