# TDT4173 Machine Learning Course Project

September 22, 2023

This is the official grading document. Please follow the rules here in case there are different ones in other course materials.

## 1 Submission

- Most project groups contain 3 students
- Each group submits three things:
  - Select 2 predictions on Kaggle
  - A short Jupyter notebook
    - It contains only the necessary steps to reproduce your final predictions.
    - We may re-run your notebook to check whether the results are reproducible.
    - Naming: Short_notebook_<1 or 2>.ipynb
  - A long Jupyter notebook
    - It contains all attempts in your group work e.g., exploratory data analysis, all models/algorithms no matter they are helpful or not, model interpretation
- You should submit the predictions to Kaggle and the notebooks to Blackboard
- You can use whatever languages, tools, platforms, software libraries, or file formats you want during development, but you should use Jupyter notebooks for your submission.
- If we cannot find your predictions or notebook(s) after the deadline, we will ask you to submit them again and treat them as late submissions.
- The project deadline is 12/11/2023 at 22:00 Oslo time.
  - The deadline is strict. Being late for even one minute is treated as a late submission.
  - You should submit both notebooks and predictions before the project deadline. Being late in either part is treated as a late submission.
- Late submission deadline is 15/11/2023 at 22:00 Oslo time.
  - A late submission will cause a -10 deduction in your course points.
  - If you submit even later, the whole group fails the course.
- The competition will be closed after the deadline
  - Late submitted predictions must be sent to zhirong.yang@ntnu.no.
  - We will manually calculate the ranking of late predictions.
  - Only one late submission is allowed.
- At the beginning of each notebook, write your full names, student ids, and Kaggle team name.

## 2 Grading

- Project points = base points + possible deductions
- Convert points to letter grade
  - A: 89-100
  - B: 77-88
  - C: 65-76
  - D: 53-64
  - E: 41-52
  - F: 0-40
- Base points
  - proportional to the number of Virtual Teams (VTs) you defeat
  - max 100 (defeat all VTs) and min 41 (defeat 1 VT)
  - if you defeat 0 VT, you fail the course
- Virtual Teams (VTs)
  - VTs are prepared by the teachers and assistants
  - A VT uses the same data as you have
  - All VTs will be frozen after 20/October/2023
- We use the private leaderboard to calculate base points
  - You can see only the public leaderboard before the competition closes
  - You select two predictions before the deadline (your short notebooks should be able to reproduce the selected predictions)
  - We use the best between your selected in the private leaderboard to calculate your base points
- More about the reproducibility
  - No external data is allowed.
  - Your short notebooks will be run in an offline setting
  - You can use a few programming constants; but writing massive data in code is not allowed.
  - Your short notebooks should start from the given raw data
    - Before running your code, we will delete all your non-code files.
    - During the running, your program may still store some middle results as disk files.
  - You can directly specify a few hyperparameter values in the short notebooks, but you must describe how you obtain the values in the long notebook.
- Possible deductions
  - pass the individual assignment in the second chance (-5); the deduction is for individual team members.
  - late submission ($\leq$3 days after the project deadline) of the project (-10)
  - no exploratory data analysis (-3). To avoid the deduction, you should do at least four or more items of the following list:
    - Search domain knowledge
    - Check if the data is intuitive

    - Understand how the data was generated
    - Explore individual features
    - Explore pairs and groups of features
    - Clean up features
  - only one type of predictor is used (-3). To avoid the deduction, you should try two or more types of predictors in the long notebook (e.g., XGBoost and Random Forest). It is allowed to use only one type of predictor in the short notebook and Kaggle predictions.
  - no feature engineering (-3). To avoid the deduction, you should try one or more feature engineering techniques (i.e., feature selection and/or feature extraction) in the long notebook.
  - no model interpretation (-3). To avoid the deduction, you should use one or more model interpretation techniques (e.g., PDP, feature importance, LIME) in the long notebook.

- All deductions are binary. That is, there are only two possibilities: full deduction and no deduction. There is no intermediate deduction.

- A grading example: a student
  - passes the individual assignment in the first chance
  - submits the project in time
  - his or her team defeats 7 of 10 VTs
  - no model interpretation in the notebooks (-3)
  - Then the student's course points
    - base points $= 41 + \frac{100-41}{10-1} \times (7-1) \approx 80$
    - course points $= 80 - 3 = 77$
    - rounded letter grade is B