Project Proposal for ML Practical

Your Machine Learning project starts with a proposal. On Brightspace we offer a few topics that you might work on, or you might propose your own topic. This proposal assignment assesses whether you're able to deliver a clear and complete description of the project you intend to do. **It is strongly recommended that you discuss your project with your TA**. They can help you figure out what is realistic for your project.

Note that this is not a "just fill it out" assignment. You will need to research the topic, dive into the data and look at what previous ML researchers have done before you. Since this all builds on top of eachother you **cannot assign each topic to a different person**. Writing a good proposal takes time, but it will save you that same time in the future. The corners you cut here will cut your fingers in the future.

The numbered questions offer a series of things you need to answer. We will offer an example solution on Brightspace. Please focus on adding **reasoning for your choices and your answers. This is required for every question.** The sub-questions for each item give you some things to think about, but are not a to-do list. You should be able to do the whole assignment in max. 2 pages of text (excluding figures and tables and oversized headers), though this will not be strictly enforced.
Add \usepackage[a4paper]{geometry} to get reasonable margins if using LaTeX..

For this assignment you will be assessed on **the proposal and the preliminary research** but not yet on the project. Questions 3-5 will be assessed on how well you reason your design choices. Proposing an amazing high-effort project with no explanation will not give you a good grade here. A very well argued trivial project can get you a good grade here. The project that you'll end up doing should be somewhere in between. Your TA will give you feedback on the level of the proposed project, but this will not be part of the grade here (but it will be graded at the poster presentations).

Please submit your proposal as a **PDF file, following the itemized list** we set out here (to make grading easier). I recommend using **Overleaf to manage your LaTeX**.
You should make this proposal in a **group of 3**. If it is impossible to form a group of 3 please get in touch with a TA.

1. Preliminary domain knowledge (1 point)

Introduce your task to a reader (who hasn't read the Brightspace topic introductions). What is the topic? What is the task that you'll be trying to solve? What is already known or believed *before* the data is investigated? Is there a specific task that you're trying to solve or a (research) question that you're trying to answer?

Required: The ML task and at least one insight before going into the data

2. Preliminary data exploration (2 points)

Which features are in your dataset? How are your features distributed? Are some of your features correlated with each other or with the target label? How many classes are there and are they balanced? Does some of the information here contradict or confirm prior domain knowledge? Are there any missing values, or outliers in your data?

Here you should include some plots that show what your dataset is all about. You might use histograms, scatter plots, or whatever helps you understand your data. Please show at least **one dimensionality reduced visualization (PCA, t-SNE, MDS),** as well as **two other plots** that help explain your data. Don't forget to **interpret your plots.**

Required: 1 dimensionality reduced plot, 2 other plots

3. Proposed preprocessing (1.5 points)

What preprocessing steps will you be doing? Topics that you might cover are: normalization, feature extraction, dimensionality reduction / feature selection, value imputation (for your NaNs or other bad data), splitting your data, dealing with class imbalance or something very specific to your topic.

You don't need to explain the details of how these methods work, just which methods you'll be using and **why** (where possible, connect this to your findings from Questions 1 & 2). Make sure it is very clear to the reader what will happen. I recommend a draw.io flowchart to show what happens.

Required: the proposed preprocessing steps
Recommended: 1 draw.io flowchart that shows the preprocessing steps and their order

4. Proposed model(s) + baseline (1.5 points)

Start with your baseline model. Which simple model will be you comparing your solution to? (Note that a baseline may also have simplified preprocessing).
Which “proper” model(s) will you be using to get a well performing ML system? Explain your reasoning. How will you be training your model(s)? What hyperparameter tuning strategy will you implement? Which hyperparameters will you tune? You may assume that the reader knows about common ML methods, so you don’t need to explain e.g. what an SVM is, but you should explain your choice of kernel function.

Required: At least 1 baseline and at least 1 “full” model

5. Proposed evaluation (1.5 points)

How will you assess your model(s) in the end? Here you can think about metrics like accuracy, F1, AUROC, MAE, MSE, etc., but you can also discuss things to reflect your system’s behavior, such as confusion matrices. I encourage you to also evaluate beyond predictive performance. Consider train cost, inference cost, model size, social biases, …

Required: At least one quantitative metric for predictive performance,
Required: At least one other evaluation

6. Model usage (1 point)

How, why and by whom could your model be used? What needs to happen so users can make a “request” and get a usable “response”? You may cover preprocessing, but also how you turn your model’s prediction into a usable answer for the user. Simple things like turning “class 1” back into “Iris Versicolor” should be implemented by you. Larger things like streaming EEG data should be considered, but not implemented.

Required: At least one “user group” and one “use case”
Required: something that happens “before” and something “after” your model

7. Risk assessment (1 point)

What are the potential risks related to your project? What issues may you expect during development? How would you circumvent this? How likely is this project to be successful? What are societal/user risks when your model is deployed publically?

Required: At least one “threat” to your project development and an estimate of success.
Required: A consideration of possible risks of having your model deployed.

8. Individual learning outcomes (0.5 points)

Ocasys (and the first lecture) define several learning outcomes that you will achieve with this course. What things would you like to learn beyond this, and why? Each of you should define something *you* individually are planning to learn.

Required: one learning outcome (full sentence) *per group member,* with reasoning

Some suggestions: project management & steering, IDEs, writing tests, Transformers, clean code, …

9. Member contributions (no points)

What did each member contribute to this proposal? This question is not graded, but should create a “paper trail” in case there’s conflict later.

During the course you might find that you end up deviating from the proposed project. This is normal, but don’t lose sight of your plan/task. If you’re not sure about the changes, discuss with your TA whether your changes are good.