# Final assignment

*Data Science course*

## Goal of the assignment

For this assignment, you will write a short research paper on a series of machine learning experiments with Kaggle data.

## Submission guidelines

- **Submit your document to Brightspace**,
- Please submit your report as a single pdf (not zipped), together with your code (as separate file), in Brightspace. Don't list your code in the report.
- Work in teams of 2 (there will be a plagiarism check)
- **Deadline: May 22.**
- The length of the paper should be 6 to 8 pages (8 is the maximum). Keep it concise. Writing short paper is an important exercise.
- I advise you to write your report in Overleaf. This is a template that you might use: https://www.overleaf.com/latex/templates/cs330-assignment-template/wjtyxcpwpfmw
  If you need more space, you could opt for a two-column template: https://www.overleaf.com/latex/templates/acm-conference-proceedings-master-template/pnrfvrrdbfwt

## Intermediate deadlines

The final deadline is May 22.

To ensure everyone is on track, we have two intermediate reporting dates:

- April 24: you submit a draft of your introduction with your tentative research question(s)
- May 8: you submit a draft of your method section

## Assignment instructions

### Development and experimentation

You are going to work with the Home Depot Data https://www.kaggle.com/c/home-depot-product-search-relevance/data , but as opposed to assignment 2 you are going to develop your own method.

You work in Python and build your implementation upon existing packages and models (sklearn, tensorflow, huggingface, spacy, etc., or something else).

You can use the Kaggle page to get inspiration from the notebooks there, but you write your own code.

You run a sequence of experiments to evaluate your method.

## Paper writing

Your paper should have the following structure. You can re-use parts of your report for assignment 2 and take advantage of the feedback you received on your report.

1. **Introduction**, consisting of:
   Context, problem description, task definition, research question(s).
   You can specify your own research questions: you might focus on a comparison between models, methods, or features, or ask a specific question relating to the data at hand (e.g. which are easy and difficult items to classify, or what the effect of specific pre-processing is on the data).
   A separate section with background literature is not necessary.
2. **Data**
   Description of the data; you could include (parts of) the data exploration from assignment 2.
3. **Methods**
   Describe what you did. This includes: a description the data pre-processing, the models you used, the hyperparameter organization and the evaluation procedure (train/test splits, metrics used). Don't go into details about the models and packages you used.
4. **Results**
   Tables and description. Focus on the interesting results and discuss them. Also compare your results to the results from assignment 2. Additional analyses of methods or features are welcomed. Use explainable AI techniques to analyse miss classified samples for example.
5. **Conclusion**
   Answer your research question(s) and discuss implications of the results.