

---

## Project in Introduction to Machine Learning

To be solved in Groups of 2 or 3 Members

Submit in moodle by 31th December, 2024, 23h59

You need to deliver:

- project proposal**: 1-page pdf, due by **Nov 15** (10%)
- send the code of your work**: zip file, due by **Dec 31** (15%)
- write a Report**: 6 pages most in double column, pdf format, due by **Dec 31** (50%)
- do an oral presentation**, 15 min most, date to be defined (in **January**) (25%)

**Description**: here you have the opportunity to explore the application of machine learning techniques in the context of a real-world data set. The project ideas and data are presented in the annex.

### Project Proposal:

You must turn in a brief project proposal (1-page maximum). Read the list of available data sets and potential project ideas. We suggest projects on data that has already been collected, so try to work on existing data sets (exceptions are allowed). Note that even though you can use data sets you have used before, you cannot use as a class project something that you started doing prior to the class.

Although you can use deep learning, you should not use only deep learning. You can compare conventional machine learning methods with deep learning approaches, you can use deep learning to extract features used afterwards by a conventional method, etc.

**Project proposal format**: Proposals should be one-page maximum. Include the following information:

- Project title.
- Team.
- Data set.
- Project idea. This should be approximately two paragraphs.
- Software you will need to write.
- Papers to read. Include 1-3 relevant papers. You will probably want to read at least one of them before submitting your proposal.

### Annex

Potential project ideas:

1. **Frame the project within your PhD thesis.**
2. Participate in (or use data from) international competitions
  - \*Kaggle Competitions (<https://www.kaggle.com/competitions>)
  - \*DrivenData (<https://www.drivendata.org/competitions/>)
  - \* [http://grand-challenge.org/All\\_Challenges/](http://grand-challenge.org/All_Challenges/)
3. **Predicting Kinase Inhibitor Activity** with the KinDEL Dataset: see corresponding pdf document in the zip file.
4. **"Relative or Absolute Decisions?"** In here, you will work on an ordinal classification problem. But instead of predicting the output for an observation (for instance in "Excellent", "Good", "Fair", "Poor"), one can also tackle it as a relative decision: a) design a model to receive a pair of observations and predict the best one; b) use that model to infer the absolute class of an observation. In this project, you should compare the conventional and the proposed approaches in a dataset of your choice.

5. **"The output as an input"**, Instead of learning a model to predict the output from the input, in here we will try to predict how well a certain output candidate agrees with the input. For a certain <input, output candidate answer> the model should predict the 'quality' the output candidate for the input observation. You can select a dataset of your choice (binary or ordinal classification problem). For instance, the red wine quality dataset used in the second assignment (<https://archive.ics.uci.edu/dataset/186/wine+quality>).