# CS461 Machine Learning
# Course Project Guidelines

**Overview:**

Your class project is an opportunity for you to explore an interesting machine learning problem of your choice in the context of a real-world data set. The purpose of the project must be related to some aspect of the material but may explore an avenue that was left unaddressed in class. The projects can be done individually, or in teams of two students. Your project will be worth **20%** of your final class grade, and will have 4 deliverables:

1. **Proposal**: (Feb 4th) at most one-page, 12-point font, single spacing, 1 inch margins (10%)
2. **Midway Report**: (Feb 25th) 4-5 pages (25%)
3. **Poster Presentation**: (Mar 21th) (20%)
4. **Final Report**: (Mar 21th) (45%)

---

## Project Proposal (Due Date: Sunday, 4th, 23:59 PM) (10%)

In order to help guide your choice of a project, you are required to submit a brief proposal (at most one-page, 12-point font, single spacing, 1 inch margins) that describes the idea for a project. The proposal should be one page maximum. Include the following information:

- Project title
- Project idea. This should be approximately two paragraphs.It should identify the project type, the problem you plan to address, the motivation for why you find the problem important or interesting, any previous work you already know about (if applicable), and a rough tentative approach to solving the problem .
- Data set. Describe the dataset, such as how many columns, rows, the source of the dataset, etc.
- Software you will need to write.
- Teammate (if any) and work division. I expect projects done in a group to be more substantial than projects done individually.
- Midterm milestone: What will you complete by February 25th? Experimental results of some kind are expected here.
- Page limit: 1 Page

Note: If you are having trouble writing a proposal, feel free to contact the instructor (Dr. Ali Aburas).

---

## Midway Report (Feb 25th) 4-5 pages (25%)

This should be a 2-4 pages short report, and it serves as a check-point. It should consist of the same sections as your **final report** (background, method, experiment, conclusion and references). The introduction and related work sections should be in their almost final form; the section on the proposed method should be almost finished; the sections on the experiments and conclusions will have whatever results you have obtained, as well as “place-holders” for additional results you plan/hope to obtain.

Grading scheme for the project report:

- 70% for proposed method  and some experiments results so far
- 30% for the design of upcoming experiments

---

## **Poster Presentation (Mar** 21th**)** (20%)

All project members should present during the presentation hours. The session will be open to everybody (**if applicable**).

You can create a bunch of "normal" presentation slides, print out each one on a piece of (letter-sized) paper, and put them all together on a poster board. I will provide a template if you need it.

---

## **Final Report**: (**Mar** 21th) (45%)

The final report should include about four (3-4) pages of text per person (not including figures) in the same format as the **Proposal** and **Midway Report.** The final report includes sections such as (background, method, experiment, conclusion and references).

---

## Project type/Suggested Projects

There are various types of projects you can consider:

1. ***Applying techniques you have learned***: The project may be very practical in terms of applying techniques you have learned in the course to a real problem such as classification of email messages.

2. The project may involve **designing or adapting existing algorithms** to a novel class of problems. For example, how might we solve multiple related classification tasks?
3. ***Comparison of algorithms***: Throughout the course, we've been discussing various algorithms and their properties. Oftentimes, algorithms don't work like expected and algorithms may need to be adapted or modified to better fit the assumptions inherent in the data. What work needs to be done to adapt a model to an interesting set of data that you've found? How do various algorithms perform on the same set of data?
4. ***Missing information***: Various real world classification problems involve missing components in the input vectors. How can you deal with such missing information? Do you expect your method to degrade rapidly if more information is missing?
5. ***The Hyperparameter Selection (e.g., kernel function in SVMs)***: Machine learning algorithms automatically adjust (learn) their internal parameters based on data. However, there is a subset of parameters that is not learned and that has to be manually configured. The performance of a model can depend on the choice of its hyperparameters. Four commonly used optimization strategies: Grid search, Random search, Hill climbing, and Bayesian optimization.
6. **Approaches to Solve Imbalanced Classes**. Imbalanced data typically refers to a problem with classification problems where the classes are not represented equally. There are systematic algorithms that you can use to generate synthetic samples. A simple way to generate synthetic samples is to randomly sample the attributes from instances in the minority class. It consists of removing samples from the majority class (**under-sampling**) and/or adding more examples from the minority class (**over-sampling**). The most popular of such algorithms is called **SMOTE** or the Synthetic Minority Over-sampling Technique.

**Note**: You shouldn't worry about getting "great" results. The idea and your understanding of the machine learning issues involved are much more important than getting "great" results.

---

## Some data repositories you might find useful:

the UC Irvine Machine Learning Repository — https://archive.ics.uci.edu/
KDD Repository (Various) — http://kdd.ics.uci.edu/
Protein data bank (Genome) — https://www.rcsb.org/
Protein structural database (Genome) — http://scop.mrc-lmb.cam.ac.uk/scop/
Cancer classification data (Medical) — https://www.kaggle.com/code/shubhankartiwari/cancer-classification

Newsgroups (Text) — http://www.ai.mit.edu/people/jrennie/20 newsgroups/
Reuters Documents (Text) — https://kdd.ics.uci.edu/databases/reuters21578/reuters21578.html
4 Universities (Text) — http://www.cs.cmu.edu/afs/cs.cmu.edu/project/theo-20/www/data/