# Advanced Learning and Data Analysis

## Project Reports Guidelines - Spring 2023

## Course Project Guidelines

Your class project is an opportunity for you to explore an interesting data mining problem of your choice in the context of a real-world data set. Below, you will find some project ideas, but the best idea would be to combine what we have learned in this course with your own research interests. Your class project must be about new things you have done this semester; you can't use results you have developed in previous semesters or in other classes.

Projects can be done in teams of three to four students or individually. Group members are responsible for dividing up the work equally and making sure that each member contributes. For each project, the TAs will be your project consultant/mentor. If you have questions, please consult with the TAs before finalizing the project proposal. The final responsibility to define and execute an interesting piece of work is yours. You are strongly urged to consult the TAs or the instructor early on if your project will rely purely on simulated data or if you intend to do a learning theory related project.

**Deliverables**: Your Project Report will have 3 deliverables (see the course schedule for deadlines):

1. Proposal, 1 page (8%)
2. Midterm Report (16%)
3. Project Final Report and Code (76%)

**Format**: Your paper should use the 2-column ACM Conference Paper Template. We highly recommend using LaTeX (use sample-sigconf.tex), but you may also use the Word template. Your paper should look something like this. The page limits for each deliverable (see sections below) are strict! We will not consider text past these limits.

- **Group Number and Section**: Put your course section (CSC 422 vs CSC 522) and your group number somewhere in the title of your document.
- **Collaboration**: If you are collaborating in LaTeX, you should use Overleaf (NCSU has a license), which even provides a template for the ACM format (make sure to use the 2-column sample-sigconf.tex file).
- **Citations**: Per the ACM format, you should use short in-text citations [1], which correspond to full citations at the end of the paper. This can be accomplished with LaTeX's \cite command. See the example paper for reference.

Each deliverable of your project will be evaluated based on several factors:

- **Novelty**: The novelty of the project ideas and applications. The groups are encouraged to come up with original ideas and novel applications for the projects. A project with new ideas (algorithms, methods, theory) on data mining or new, interesting applications of existing algorithms is scored higher than a project without much new idea/application. You should *explicitly* note your project's novel aspects in your proposal and reports.
- **Experiments**: The extensiveness of the study and experiments. A project that produces a more intelligent system by combining several ML techniques together, or a project that involves extensive experiments and thorough analysis of the experimental results, or a project that nicely incorporates various real world applications, are scored higher.
- **Writing**: The writing style and the clarity of the written paper.

# Project Proposal (1-2 pages)

You must include the following information (see evaluation criteria; no format requirements):

1. Project title
2. Dataset description (see below).
3. Project idea. This should be approximately two paragraphs and show how your project is novel (see below).
4. Software you will need to write. Your project must involve some implementation.
5. Papers to read: Include 1-3 relevant papers. You will probably want to read at least one of them before submitting your proposal.
6. Teammates and work division. Each teammate should have clearly defined roles and aspects of the midterm milestone that they are responsible for.
7. Midterm milestone: What will you complete by Midway report? Experimental results of some kind are expected here.

**Datasets**: A list of suggested projects and data sets are given below. You are encouraged to use one of the suggested data sets, since we know that they have been successfully used for data mining and machine learning in the past. If you prefer to use a different data set, you must have access to this data already and present a clear proposal for what you would do with it.

## Novelty

An important feature of your project, which your proposal must convey, is novelty. It is important that your project go beyond what you have done the homeworks. Here are some examples of how you might do that, and your project may have more than 1 novel aspect:

- Collect a novel dataset (e.g. using web scraping) or transform a dataset that is not immediate appropriate for machine learning tasks (i.e. unstructured data, text data, time-series data)
- Use feature engineering based on domain knowledge, going beyond the already supplied features in the dataset (not just PCA or feature selection, but some novel approach)
- Implement a novel/complex machine learning technique from a research paper (where the code is not already provided)
- Implement a known machine learning technique and modify it in some way to address a specific challenge of your dataset (note this does not have to be a perfect idea, as long as it is reasonable and feasible)
- Use your dataset to explore a meaningful hypothesis where: a) the answer is non-obvious and b) the answer goes beyond "yes" or "no". For example, in addition to predicting whether a flight will be delayed from a dataset, also answer the question: "what factors contribute to flights being delayed?"
- Create new knowledge through analytics and pattern-finding. For example, in an educational dataset, after clustering students by their attributes, try to make sense of the

clusters you found. What patterns of behavior does each group engage in? What might this mean?

### Common Mistakes for the Proposal

- **Over-promising**: Some projects can be quite ambitious, which is great, but make sure you have time and a concrete plan to carry out all of your proposed work, e.g. through a detailed timeline.
- **Vagueness**: Some projects can be vague about what you would be doing to solve your problem (e.g. what learning, feature engineering, evaluation, etc. approaches). Concrete details help you and also help us give you better feedback.
- **Just using a library**: Your project needs some elements of novelty (one for 422 and two for 522). Examples are in the project description. You should not simply by using an existing library with an existing Kaggle dataset to train and test a classifier. See the Novelty suggestions of how to improve projects.
- **No justification of choices**: This will be more important in the mid-way report, but make sure your choices of which classifiers, pre-processing approaches, evaluation metrics, etc. are well-justified. Why did you choose this one out of all the possibilities?

## Midway Report (4-5 pages)

This should be a 4-5 pages short report using the ACM template. It serves as a checkpoint for instructors to give you formative feedback. It should consist of the same sections as your final report (see here for descriptions), with a few sections in progress:

- **Introduction** and **Background** sections should be almost in their final form
- **Proposed Method** should be mostly finished
- **Experiment** and **Results** will have whatever results you have obtained, as well as "place-holders" for the results you plan/hope to obtain.

**Meeting Schedule**: To facilitate collaboration, your report must include a schedule of **at least 4** meetings that your group will hold. Include specific dates/times/durations for each meeting, which **all members agree** that they can make (though these can be rescheduled later if needed). Your final report should include a report of who attended each meeting. This does not count towards your page limit.

**Evaluation Rubric**: The rubric for each section can be found on Moodle by clicking on the assignment.

**Page limit**: The midway report should be 4-5 pages max plus an optional appendix (below).

**Appendix**: If you have additional figures and tables that you would like to include but are concerned about length, you can include up to 2 pages of an Appendix. It should include only

figures and tables, with short (1-2 sentence) captions. You should still put the most important figures/tables in your results section with the part of your paper that discusses them.

# Final Report (6-7 pages) and Source Code

The main objective for the final report paper is to let you have the opportunity to do independent research-oriented work by reading and analyzing research papers, surveying current frontiers of some research topics, and experimenting with your own new ideas on data mining.

Your final project report should clearly articulate the rationale behind your project, in addition to simply reporting what you did and the results that were achieved. It should roughly have the following format (see Section descriptions below):

1. Introduction and Background
    a. Problem statement
    b. Related work
2. Method
    a. Novel Aspect(s): What makes your project work novel?
    b. Rationale: Why should this method perform well, and better than other reasonable baseline methods?
    c. Approach: Description of its algorithms
3. Plan & Experiment
    a. Dataset(s): Describe your dataset; number of features, number of rows,missing data, pre-processing procedure if any, and so on
    b. Hypothesis: What do you expect your results to be?
    c. Experimental Design: Details of the experiments
4. Results
    a. Results: Description of results
    b. Discussion: Help to interpret the results and relate them to prior work
5. Conclusions

**Submission: Source Code & Dataset:** When submitting your project on Moodle, you should submit two separate files:

1. Your report as a PDF file. Please name your final report in the way of 'P+ProjectID_firstKeyWord_secondKeyword', like P50_data mining_data analysis.
2. A .zip file including the **source code and dataset** for your project, including a README with setup instructions. Your code will not be directly graded, but will be used to confirm the claims in your report.
    - You should additionally include a link to your GitHub repository in your final report, but please include your source code in your submission.
    - Also include either 1) the dataset for your project (if <1GB), or 2) a link to download the dataset.

**Page limit**: The midway report should be 6-7 pages max plus an optional appendix (same rules as the midway report).

**Evaluation Rubric**: The rubric for each section can be found on Moodle by clicking on the assignment. More broadly, the final report paper will be evaluated based on the following general factors:

1. The novelty of the paper. Final report papers reporting the student's own new ideas are favored in comparison to papers analyzing existing works.
2. The depth of the understanding and the critique on existing works. Papers involving good analysis and critique of existing works are favored in comparison to papers that only describe the existing works.
3. The writing style and the clarity of the written paper.

## Section Descriptions

Below we give advice on how to write each of the paper sections (you can also refer to the rubric above for how they will be evaluated):

**1. Background**

- Problem: As clearly as possible, describe the general problem you are working on and explain the specific refinements and special cases that you addressed.
- Related Work: Briefly survey the existing work that has been done on your problem, as well as the existing work that has been done on the approach(es) you are considering. You should cite some relevant references (at least 5, not including textbooks). Your survey need not be exhaustive, but you should try to cover the most important prior work if you can.

**2. Methods**

- Novel Aspect(s): You should explicitly describe what makes your project novel (see the Proposal section for the novelty requirements). This can point to later sections where more details are given.
- Approach: As clearly as possible, describe the machine learning technique(s) that you applied to the problem, and clearly specify the final system(s) that were implemented. This section should describe your approach in a general way, such that others could implement it on a new dataset. In your Experiment section (described later), you will give the specific details of how you applied your approach to your own dataset.
- Rationale: Explain why you implemented the systems that you did. Specifically, explain the simplifications you made along the way, and why you made these particular choices (and not others). For example, you might detail other approaches that could be applied to the problem and explain why these were not pursued.
    - For example: "We chose Lasso regression, rather than a linear regression, because we are interested in identifying the features which are most relevant for predicting our outcome variable."

**3. Plan & Experiment**

The most important (and most difficult) part of any research project is figuring out how to evaluate the results. You should state concrete goals for your project. In particular, you should identify at least one concrete hypothesis that you think would be interesting to test with your implemented system and outline a specific plan for experiments that you would need to conduct in order to test each hypothesis. It should be stated as the concrete goals for your project. You should include the following information:

- Datasets: describe all the relevant information of the dataset that we need to know to understand your project.
- Hypotheses: Clearly state the main questions that you investigated in this project. These should be identified before you run the experiments! Ideally, these should be interesting questions whose answer is not obvious beforehand, where the answer would be interesting no matter how the experiments turned out.
- Experimental Design: Once you have settled on some good questions, it is important to figure out how to answer them. For each main question, describe a series of experiments that are designed to answer the question. Explain the difficulties faced in designing these tests and explain why your experiments will overcome these difficulties to yield a definitive answer.
    - See the checklist below for things to be sure to include.
- *Note*: You should not phrase your project goals as "my proposed approach will work great at solving problem X". Rather, you should phrase your goals as tests of specific hypotheses whose outcome would be interesting no matter what happened (so that your project would succeed however the tests turned out).

Checklist of things to include in experiment description:

- ✓ How you split your data into training/validation/test, or your CV approach.
- ✓ How you tuned hyperparameters (e.g. CV), and/or values you used.
- ✓ Which evaluation metrics you used.
- ✓ Which and baselines you compared your model against.
- ✓ Anything else we need to know to replicate your experiment.

**4. Results**

- Results: As clearly as possible, describe the results you obtained with your tests. Use plots, graphs, and tables, as necessary. Make sure it is easy to understand what happened.
- Discussion: In the end it is important to critically evaluate the results of the experiments. Did the experimental results answer the questions? If so, then what were the answers? If not, then why did the results fail to yield definitive answers? Is it then possible to formulate a new experimental strategy? How do these results relate to the prior work you discussed earlier?
- *Note*: It is important to demonstrate critical thought in this part of your assessment. You will be graded more on the strength of your reasoning rather than how the tests actually

turned out. For example, it is perfectly acceptable if you did not achieve definitive answers to your questions, as long as you can recognize this, explain why it happened, and suggest additional tests that might yield more definitive results.

**5. Conclusions**

- Lessons Learned: What specific things did you learn from doing this project? Did you learn anything about the problem itself, the approaches you tried, or the experiments you conducted? If there are any good ideas you came up with in the end but did not have time to pursue, this would be a good place to mention them.

**6. Meeting attendance**: Your final report should include a report of who attended each meeting. This does not count towards your page limit.

## Requirements for CSC 422 vs. CSC 522

CSC422 and CSC522 have the same required sections for their project. However, the requirements for novelty will be different. For CSC 522, we expect that each project will be truly novel (e.g. at least 2 of the Novelty elements listed above). For CSC 422, less novelty is expected (e.g. a single novel aspect). Since the difficulty of the proposed problem will shape the entire project, evaluation of each project milestone will consider these differences.

# Dataset Suggestions

Here are some project ideas and datasets for this year:

## Dataset Repositories

These contain many datasets:

- Kaggle contains datasets with specific analysis challenges: Kaggle.com
- The UCI machine learning repository: https://archive.ics.uci.edu/ml/index.php
- CORGIS datasets: https://corgis-edu.github.io/corgis/
- The PSLC Datashop: Contains educational datasets: https://pslcdatashop.web.cmu.edu/

## EcoNET Dataset

This semester you will have the opportunity to work with a client and an original dataset and problem. The client is the NC Climate Office and their data comes from the EcoNET project. The project collects data from 50+ weather stations across North Carolina every minute. Sometimes these measurements are flagged by an automated Quality Assurance (QA), and they must be manually reviewed for errors. Your goal is to build a model that can predict which flagged measurements will turn out to be true errors, and which are false positives. The dataset and documentation can be found here.

## NBA statistics data

http://www.cs.cmu.edu/~awm/10701/project/databasebasketball2.0.zip
Contains 2004-2005 NBA and ABA stats for:

- Player regular season stats
- Player regular season career totals
- Player playoff stats
- Player playoff career totals
- Player all-star game stats
- Team regular season stats
- Complete draft history
- coaches season.txt: nba coaching records by season
- coaches career.txt: nba career coaching records
- Currently all of the regular season

Example project ideas:

- Outlier detection on the players; find out who are the outstanding players.
- Predict game outcomes.

## Netflix Prize Dataset

The Netflix Prize data set gives 100 million records of the form "user X rated movie Y a 4.0 on 2/12/05." The data is available:

- http://www.netflixprize.com/
- https://www.cs.uic.edu/~liub/Netflix-KDD-Cup-2007.html

Project ideas:

- Can you predict the rating a user will give on a movie from the movies that user has rated in the past, as well as the ratings similar users have given similar movies?
- Can you discover clusters of similar movies or users?
- Can you predict which users rated which movies in 2006? In other words, your task is to predict the probability that each pair was rated in 2006. Note that the actual rating is irrelevant, and we just want whether the movie was rated by that user sometime in 2006. The date in 2006 when the rating was given is also irrelevant. The test data can be found at: http://www.cs.uic.edu/%7Eliub/Netflix-KDD-Cup-2007.html#download

## Enron Email Dataset

The Enron Email dataset contains about 500,000 emails from about 150 users. The data set is available here: http://www.cs.cmu.edu/%7Eenron/

Project ideas:

- Can you classify the text of an e-mail message to decide who sent it?

## The CSEDM Data Challenge

An educational dataset challenge. The goal is to predict whether students will succeed on their next programming problem, based on past attempts.

Challenge Website: https://github.com/thomaswp/CSEDM2019-Data-Challenge

## Precipitation data

This dataset has includes 45 years of daily precipitation data from the Northwest of the US:
Download Dataset: http://www.jisao.washington.edu/data_sets/widmann/

Project ideas:

- Weather prediction: Learn a probabilistic model to predict rain levels.
- Sensor selection: Where should you place sensor to best predict rain?

## Image Segmentation Dataset

The goal is to segment images in a meaningful way. Berkeley collected three hundred images and paid students to hand-segment each one (usually each image has multiple hand-segmentations). Two-hundred of these images are training images, and the remaining 100 are test images. The dataset includes code for reading the images and ground-truth labels, computing the benchmark scores, and some other utility functions. It also includes code for a segmentation example. This dataset is new and the problem unsolved, so there is a chance that you could come up with the leading algorithm for your project.

Download Dataset: http://www.cs.berkeley.edu/projects/vision/grouping/segbench/

## WebKB

This dataset contains web pages from 4 universities, labeled with whether they are professor, student, project, or other pages.

Download Dataset: http://www-2.cs.cmu.edu/~webkb/

Project ideas:

- Learning classifiers to predict the type of webpage from the text.
- Can you improve accuracy by exploiting correlations between pages that point to each other using graphical models?

## Email Annotation

The datasets provided below are sets of emails. The goal is to identify which parts of the email refer to a person's name. This task is an example of the general problem area of Information Extraction.

Download Dataset: http://www.cs.cmu.edu/~einat/datasets.html
Project Ideas:

- Model the task as a Sequential Labeling problem, where each email is a sequence of tokens, and each token can have either a label of "person-name" or "not-a-person-name".

## Object Recognition

The Caltech 256 dataset contains images of 256 object categories taken at varying orientations, varying lighting conditions, and with different backgrounds.
Download Dataset: http://www.vision.caltech.edu/Image_Datasets/Caltech256/
Project ideas:

- You can try to create an object recognition system which can identify which object category is the best match for a given test image.
- Apply clustering to learn object categories without supervision.