

Team Project

For the project, you will need to submit only one report per group. You are encouraged to choose a project related to your own research interests, and please feel free to discuss your project with the instructor if you want. To avoid confusions to the instructor, it is strongly recommended that the team member who submitted the proposal will also submit the presentation file and final written report.

Be aware of the following requirements and deadlines (One submission per team):

1. **The Project Proposal** is due at Canvas at **11:59pm, March 22, 2021**. The purpose of the proposal is to get you started, and also allows the instructor to provide feedback to your project. It shall be 1 ~ 3 pages. You will need to provide the following information:

- (a) Your name(s)
- (b) Project description
- (c) How and where you obtained the data
- (d) Scientific Research questions you may want to address and corresponding machine learning & statistical learning methods

You will receive full credit on the project proposal if you provide all these information. For the data set, you can just list a website in your proposal where I can find them. You don't need to submit your data set.

2. **The Final Project:** The students will need to submit the following two deliverables to Canvas:

- (a) The **presentation slide** (either pptx or pdf version will be fine) and **recorded oral presentation video** are due at Canvas at **5:00pm, April 25, 2021**. There are no specific guidelines on the presentations, and the common sense applies, e.g., write all team members' names somewhere on the first slide, highlight your problem, data set, main ideas/methods, and conclusions. Ideally you or your team should prepare for the slides so that each teammate member can present about 3.5 minutes, e.g., a team of 2 students will prepare for slides for 7-minutes presentation, and a team of 3 students should prepare for slides for 13-minutes presentation, etc.
- (b) The **final written report** of your team project is due at Canvas at **5:00pm, April 28, 2021**. In your writeups, we expect clear explanations of models chosen, hypotheses tested, and findings analogous to what you would produce for a consulting project.

Important: For the purpose of this class, at the end of conclusion section of your final report, please **add a subsection for lessons you learned** from this project or this course. You can also write any suggestions to the instructor.

Grading: The final project will have a weight of $50\% = 5\%$ (proposal) + 15% (oral presentation) + 30% (written report) in your final course grade.

Your grade on the final written project will be very subjective, and will depend on you selecting and adhering to a logical and readable format for the report; on the appropriate use of whatever data mining technique you use; on the appropriateness in the conclusions of your report; and on the readability and understandability of the report when technical material is needed.

In general, the ideal is for all team members to receive the same grade on the final project. However, individual deductions from the team's final project grade will be assessed for failing to contribute a fair and significant share to the team's project, as determined by the teammates and the instructor.

Possible Topics of Your Project

The objective of a class project is to help you gain experience with research, and to relate what you learn to real life problems which may require you learn new techniques (or develop new methods by yourself). You are expected to present the project findings during the class and submit a summary report at the end of the semester. Below are the two types of possible projects (you only need to choose one of them).

1. **Solving a real life data mining problem.** A typical report includes problem formulation, data analysis, proposed solutions, and interpretation of results. The data set can be from your own research or the public domain, see the information below. As an example, you can choose to participate a data mining competition such as the Knowledge Discovery and Data Mining (KDD) cup, see the link below for the past KDD Cup <<http://www.kdd.org/kdd-cup>>, or the KDD CUP 2017, <<http://www.kdd.org/kdd2017/>>. Another example is “2017 Data Challenge” sponsored by the Government Statistics Section of the American Statistician Associations (ASA) that analyzes the Consumer Expenditure Survey (CE) data on the Bureau of Labor Statistics website, see <<http://magazine.amstat.org/blog/2017/01/01/data-challenge-on-tap-for-jsm2017>> for the announcement and <<https://www.bls.gov/cex/pumd.htm>> for the datasets.
2. **Numerical study of data mining methods using well-known data sets in the literature.** Note that when dealing with well-known data sets, your approach needs to be substantially different from the literature, i.e., you should do more than repeating the analysis there. Some examples are
 - Compare performance of competitive data mining techniques;
 - Ask different questions or investigate new ideas of data mining methods;
 - Identify optimal parameters of specific data mining techniques;

Note that the crucial aspect of your project is to **analyze some data sets and justify your conclusions**, not using some specific statistical models or methods we discussed in class.

Datasets: You can collect the data by yourself, use the data set from your own research or the public domain. One way to find online datasets is to use the search engine such as google. The followings are some examples of online datasets (you can use google or other search engine to find more):

1. <http://kdd.ics.uci.edu/> or <http://archive.ics.uci.edu/ml/>
One example is the KDD cup 1999 data at <http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html>
More KDD cup data can be found at <http://www.kdd.org/kdd-cup>
2. <http://www.quandl.com/> (financial and economic time-series datasets)
3. Data sets from some government websites such as <<http://www.cdc.gov/surveillancepractice/data.html>> or <<http://www.ngdc.noaa.gov/stp/satellite/goes/dataaccess.html>>.
4. <http://lib.stat.cmu.edu/DASL/>
5. <http://www.kdnuggets.com/datasets/index.html> (links to more data repositories.)
6. http://www.dmoz.org/Computers/Artificial_Intelligence/Machine_Learning/Datasets/

To inspire your projects, some concrete examples can be as follows:

- analyze some data sets in some competitions, see the links < <http://www.kaggle.com/competitions>>
- derive the relationship between sleep and selected health risk behaviors, see the paper <<http://www.cdc.gov/nchs/data/hestat/sleep04-06/sleep04-06.pdf>>

To further motivate your projects and encourage you to write up a solid project report, try to think that you want to publish your project report as a paper. There are two possible kinds of data mining or statistical learning papers (you only need to choose one).

- **Application Papers:** apply standard methods to analyze some datasets, thereby answering some important questions in real-world applications such as bioinformatics, economic, finance, banking, health-care, online advertisements, manufacturing, music, natural disasters, social networks, (bio)surveillance, warehouse, logistics, etc.
- **Methodology Papers:** develop new methodologies and demonstrate their advantages as compared to the standard methods when analyzing some data sets, say, in the context of temporal data mining, spatial data mining, spatio-temporal, streaming data mining, web or graphic mining, etc.

The final written report shall **not be longer than 25 pages**, and the main body of the report is generally 5 ~ 12 pages. Only very relevant plots and tables shall be included in the body of the report, and the rest should go to Appendix. When writing up your summary report, it is useful to ask yourself the following questions: What is the work? Why is it important? What background is needed? How will the work be presented?

Here is a suggested format for your summary report.

1. **Title Page:** Project Title, author(s) (your name and email address), the submission date, course name/number;
2. **Abstract:** informative summary of the whole report (100-300 words).
3. **Introduction** includes problem description and motivation, data mining challenge(s), problem solving strategies, accomplished learning from the applications and outline of the report.
4. **Problem Statement or Data Sources:** cite the data sources, and provide a simple presentation of data to help readers understand the problem or challenge(s).
5. **Proposed Methodology:** explain (and justify) your proposed data mining strategies.
6. **Analysis and Results:** present *key findings* when executing the proposed data mining methods. For the benefit of readability, detailed results should be placed in the Appendix. Reference of computer softwares to implement your proposed data mining methods (even it is a web page) should be given.
7. **Conclusions:** Draw conclusions from your data mining practice. Unfinished or possible future work could be included (with proper explanation or justification).
**A Mandatory Subsection of “Lessons we have learned”: at the end of conclusion section, please add a subsection for lessons you or your team learned from this project or this course. Please feel free to write any comments/suggestions/remarks, or share your experiences of data mining.*
8. **Appendix:** This section only includes needed documents to support the presentation in the report. Feel free to divide it into several subsections if necessary. Do NOT dump all computer outputs unorganized here.
9. Bibliography.

Parts 3-6 constitute the main body of the paper for your primary audience. Usually, as with fictional boss in this example, your audience is intelligent but unschooled in Data Mining or Statistics. So these parts should have as little technical material as you can possibly get away with.

It is appropriate, and even recommended, to refer the reader to the appendix in part 8 if you need to provide a more technical explanation for something. Part 8 is your secondary audience - me - and should follow closely enough the "story" of parts 4 – 6 that it is easy for me to see what technical material backs up with results and discussion.

It is not necessary to number these parts 1-9 or name them as-above-mentioned. Please feel free to merge some parts or provide more informative section names if it seems natural to do so.