Course Project Guidelines
(updated 8/25/23)

This document provides guidelines for writing the course project proposal and for writing the project paper that is due toward the end of the course. The project is a significant portion of your grade, so you are expected to devote a reasonable amount of time to it and to the write-up. It is difficult to precisely quantify the total amount of time you should spend on the project and write-up—but certainly 5 hours would be **much less** than expected. Projects can be done individually or in teams; teams of two are most common but larger teams are possible with advanced approval to ensure they are sufficiently ambitious. Project papers associated with team projects should be a bit longer and more comprehensive (e.g., include a more substantial related work section).

Types of Projects

There are two main types of projects. You can do a research project, where you investigate a specific research issue. This could be original research but could also be something straightforward—such as an empirical evaluation of data mining methods or strategies for improving performance. However, based on past experience, most of you will work on an application-based project, which means that you will utilize a real-world data set and address an associated real-world problem. Data mining is largely an applied field, so even such a project, if it includes some innovation, could be considered an applied research project. Ideally you should try to do something a bit interesting—for example use a data set not previously analyzed or some variation of a previously studied problem. If possible, select a project that is of particular interest to you (e.g., relates to a hobby or special talent that you have). You should make sure that your analysis is not trivial. For example, applying a few methods to a simple data set and then doing a quick write-up might be considered too trivial. There are some items listed under the "Full Project Writeup" section that can be used to extend the project. The project can relate to any topic except if these guidelines are being used for my "Data Science for Cybersecurity" course, in which case the project must have some connection to cybersecurity, although we can interpret that quite broadly.

Project Proposal

Your project proposal must be typed, submitted electronically, and should be 1-2 pages long, single spaced. The purpose of the proposal is to make sure that you are on the right track and to give me enough information so that I can give you useful feedback. The most common issue with a project proposal is that it is unclear and ill-defined. For application-based projects, you must identify the data set in your proposal (or a set of viable data sets) since data that you need may not be available. The data mining/data science problem to be studied should be clearly identified and described. Your proposal should include the following:

- Preliminary title and the names of students working on the project (consult with me before submitting a proposal for teams of 3 or more).
- Abstract: Similar to the abstract that will ultimately appear in your paper. It should be one paragraph long and for the proposal, about 5-15 lines. It should provide a high-level summary of your project and outline your main goals.
- Brief description of what you plan to do.
    - What problem are you trying to solve? Be as clear as possible.
    - How do you formulate the problem as a data mining problem (e.g., is it classification, association rule mining, etc.)? What *exactly* are you trying to predict (for prediction tasks) and how will you evaluate your results. How will you know if your results are good? What can you compare them to? It is critical that your problem is **well-defined**.
    - What data sets do you plan to use? If you must do significant work to get the data or convert it into the proper format, then describe the process and approximate effort required. You should provide basic statistics about the data set, such as: number of examples, number of features, number of different class values, degree of class imbalance (for prediction tasks), and perhaps a list of the key features.

- What data mining packages you plan to use (e.g., WEKA. Python Scikit), realizing that for a graduate course I likely will require you to use Python Scikit. Also specify what algorithms you plan to use (e.g., decision trees, neural networks, etc.). If there are any technical challenges (e.g., class imbalance), describe how you will address them.

- It is recommended that you list a few related work papers. You should read them before going too far with the actual project. There are no firm guidelines on how large/complex a data set should be, but generally you will want thousands of examples, not hundreds, and at least a half dozen meaningful features (but ideally more than that). Identifier features and other features unlikely to be useful should not be counted.

Full Project Writeup

The final project paper should be well-written and organized and should look professional. Students completing a project for a graduate class are **required** to use the provided IEEE two-column conference template, while for the undergraduate class it is only highly recommended. The paper should be properly organized with appropriate sections and subsections with descriptive headings. Figures and tables should be included and should be readable (if the text is too small increase the text size in the figure) and have an appropriate caption. References to the figures and tables should include the figure or table number and figures should include axes labels. You should generally not just insert screen shots of results produced by a tool if it includes some information that is not useful or if the resulting font sizes are not appropriate; take the time to generate your own tables with the proper information and format them properly. Papers look best if the text is fully justified (i.e., the text should end at the right column boundary and not be "right-ragged"). Note that when academics or professionals submit a paper to a conference, they usually must follow precise formatting instructions, so a focus on format and presentation is good practice.

Below are some items that you should consider including, that are often included in the best papers. But it depends on how complex your project is and if the data requires special processing already—projects that are relatively straightforward should include as many of the items below as possible to ensure that the project is not trivial and entails more than a regular lab assignments.

- Parameter tuning: for prediction algorithms, try different settings and perhaps even visually demonstrate how the results vary. For example, results for the kNN algorithm could show how performance changes for different values of k.
- Exploratory data analysis. Visualize the distribution of key features and how they relate to the class variable. Minimally this can be done with a set of scatter plots that include the feature under consideration and the class variable.
- Consider feature selection for any methods that do not perform well with unnecessary features (e.g., kNN, linear and logistic regression).
- If your project involves building a predictive model, definitely try to explain/describe the model, not just present the results. This is easier for some models (e.g., decision trees) than others (e.g., neural networks). Most methods also will generate a measure of feature importance, so you can at least rank order the importance of the features. Then try to explain why the important features make sense—speculate on why they are predictive, possibly referring to other work that may provide the necessary background or understanding. If it is not possible to get such information for the algorithms that you focus on, you can always use some of the feature selection methods to rank the features. No paper that handles prediction should totally omit an explanation of the model or important features.

While it is important to present your results for different algorithms and compare them, understand that researchers and practitioners will not find this particularly interesting. It is worth mentioning and discussing, but there are more interesting aspects, like the explanation of the models and what features are important—or any clever or unusual steps that you take to improve results (e.g., compressing certain class values). It some cases it may even be more interesting to see how key parameters impact learning, such as the amount of training data (learning curves) or the number of trees in a random forest, or the value of k for k-NN.

The paper need not be organized exactly as described below, but it should be quite similar, since the outline below is generally what is used for most conference papers in computer science that deal with data mining or data analysis. There is no precise page or word length requirements, but what is below are rough estimates for a course-based project (not a capstone project that should be longer). As mentioned earlier, papers from a team of two or more should be somewhat longer. The key is that the paper clearly and concisely describes the project, and the project should be substantive enough so that there is a fair amount to discuss. The paper should be written for an audience who has basic knowledge of data mining and/or data science, such as those completing one course on the topic. Thus you may explain what F-measure is, but it is not required (a citation would be sufficient). If you want to include code or similar information, put that in an appendix and it should not count toward the paper length.

Length guidelines for one-person course projects (two column format holds more than one column format):

- If using IEEE 2-column single-spaced format with 10-point font: 4 – 8 pages
- If using 1-column single-spaced format with 11-point font: 5 – 12 pages
- If using 1- column 1.5 spacing format with 11-point font: 8-18 pages

Here is the suggested outline:

- Abstract: summarizes the paper and the goals of the work. It should be limited to a single paragraph and should be a maximum of 500 words. It should not provide a comprehensive summary of the paper. Rather it should motivate the problem, define it, and briefly discuss the general approach. It may or may not include some basic results, but any discussion of results should be limited to 1-2 sentences.

- Introduction: Introduces the project and what you are trying to do. Should *motivate* the problem, quickly define it and the approach taken, and may discuss some highly related work if it helps to motivate the problem or provide basic background so that the paper can be understood. Ideally you should explicitly mention the contributions of the work. Usually about ½ page to one page.

- Background: Depending on the project, you may want a separate background section, depending on how much background you want to include. For example, it may provide domain information for the domain that you are studying. If the domain is not complex, then this section may not be needed. This is generally not about related work.

- Experiment Methodology: Describes the experiments and the experiment methodology. Will describe the data sets, evaluation metrics, data mining algorithms with associated parameter settings used, the precise methodology related to the setup of experiments (partitioning of the data and k-fold cross validation), and any other details related to the experiments. There will usually be a subsection for each of the sub-topics just mentioned. If you use only well-known algorithms, you can usually just cover them all in a single paragraph (or with a citation). But if some are less well known, maybe include a paragraph on each. Results do not go into this section. Use tables where appropriate rather than long lists of items. If the data is complicated, you may want to break that into a separate section that is placed before the Methodology section.

- Results: Presents the experiment results and a discussion and analysis of the results. Normally a separate discussion section is not necessary. If there are a lot of results, try to break it down into two or three subsections if there are different types of experiments. Make good use of Tables and Figures. Figures are better than tables when you want to show something like a trend. It may make sense to have some results that are relatively low level, and then separate tables/figures to show higher level results for different experimental setups that can easily be compared. For example, you may want to summarize the results for the best set of parameters in a separate table. You should make effective use of tables and figures. Please label all figure axes precisely and provide appropriate caption names. Increase the size of the text on figures if it is much smaller than the body text. If you can fit a figure into one column of a two column formatted paper that is ideal, but if that is not possible you can use section breaks to revert to a full width column.

- Related Work: A description of related work, with citations to relevant papers. If you are doing an application paper, where you analyze some data, you are less likely to rely on a lot of related work.

Nonetheless, there almost always should be some related work discussed. If there is not that much related work to discuss, it may be possible to include the related work in the introduction (since it may provide motivation and context for the project). If there is a lot of related work, you may want to provide the bulk of it in this section but include some in the introduction too. In general, every paper should mention a minimum of 4-6 related work papers. These are not papers that are citations to relevant background (e.g., a citation to a paper on class imbalance or how decision trees work). There will almost always be related work, even if it is not a perfect match. As an example, let's say you are doing a paper on using data mining to classify Pokemon characters (several projects have done this!). You could look at other data mining papers on Pokemon, then other data mining papers on similar video games, then possible related work on e-sports. All related work papers should be cited in the paper and then should appear in the reference section. All items in the reference section should be explicitly cited in the paper.

- Conclusion: Provide your conclusion (perhaps summarize your main results). Normally will also discuss limitations and avenues for future work.
- References: Each paper should have a references section. This should include references to related work, but also references unrelated to related work. For example, if you are using Weka there should be a Weka reference (same for Python Scikit), possibly references to specific algorithms and metrics. The precise standard you utilize for references (MLA, APA, etc) is not of concern, but it should be consistent.

(sample project topics listed on next page)

Sample Data Mining Projects

Graduate Data Mining Course Projects

- Air Pressure System Failure Prediction in Scania Trucks (2019)
- Predicting Kickstarter Campaign Success (2019)
- Bike Sharing Rental Prediction (2019)
- Landscape Image Classification Using Unordered Color Data (2019)
- Spam Email Trigger Words (2019)
- Gender Prediction from OkCupid Profiles using Text Mining & Ensemble Methods (2019)
- Pokemon Type Classification (2019)
- Using Statistical Averages to Predict Preferred Roles for Overwatch League Players (2019)
- NCAA March Madness Result Prediction Model (2015)
- Authorship of Federalist Papers using SAS text mining (2009)
- A Meta-Ensemble Learning Method (2009)
    - Uses ensembles to learn faster where the data is not made available to the final classifier. A classifier is built from the output of other classifiers.
- Decision Tree Computational Complexity (2009)
    - Attempt to empirically fit the computational complexity of a decision tree algorithm.
- Using Data Mining to Target Nicotine Dependents (2009)
- Mining for Cognition (2009)
    - predicting dementia based on multiple psychological survey results
- Data Mining the College Board's Student Descriptive Questionnaire (2009)
    - Relied on cluster analysis and student had extensive domain knowledge
- Data Mining Study of Wine (2009 + more)
    - Several students have used the wine data set to predict quality of wine.
- KDD Cup 98 Donation Decision Analysis (2008)
    - This is also popular and uses the Paralyzed Veterans Administration dataset
- Using Data Mining Method to Predict Pneumonia and Nosocomial Pneumonia (2008)
- Who Uses Illicit Drugs (2008)
- Tracking and Adapting to Changes in Class Distribution Using Quantification and Semi-Supervised Learning (2008)
    - This paper was extended and published in the premier data mining conference
- Predicting University Enrollment: A Marketing Study (2007)
    - Used SAT data etc.
- Who would Enroll at Fordham University (2007)
- Mining an Electronic Payments Data Warehouse: Identifying Fraudulent Payments (2006)
- Choosing and Explaining Likely Caravan Insurance Customers (2006)
    - A COIL Challenge 2000 competition

Graduate Data Science for Cybersecurity Course Projects

- Analyzing and Projecting Factors for Detecting Malicious and Benign Websites (2023)
- Analyzing Classification Algorithms for Phishing Detection in Cybersecurity (2023)
- Single Classifier vs Ensemble-based Models for Intrusion Detection (2023)
- Edge-IIoTset: An Evaluation of Industrial Internet of Things Intrusion Detection (2023)

- Identifying Hateful Comments to Detect Cyber Threat Actors Using Machine Learning (2023)
- Experimental Comparison of Data Mining Techniques for Cloud Intrusion Detection (2023)
- A Machine Learning-Based Approach to Detect Phishing URLs in Online Advertising for Improved User Privacy (2023)

Undergraduate Data Mining Course projects

- Predicting User's Return to Website by Mining User Environment Variables
    - Had private data from a website
- Overall Car Evaluation Based On Various Specifications (2010)
- Personal Spam Filter using Decision Trees (2010)
- Predicting Favored Contraceptive Methods (2010)
- Negative Effect of Pruning under Low Recall (2008)
- Objective Empirical Comparison of Decision Tree Classifiers (2008)
- Who is Most Likely to Smoke (2007)
    - Used data collected at Fordham by the students
- Importance of Attribute Replacement to Misclassification Rate Reduction
- Comparative Study of C4.5, C5.0, and SAS Enterprise Miner
- Associative Analysis of Caffeine Intake and Lifestyle of Fordham University Students
- Efficient Market Theory versus Enterprise Miner
- Various projects related to my WISDM project from students in my research lab
- Gender and Generations: predicting Age and Income based on Social and Political Opinions (2014)
- Mining Major League Baseball (2014)
- Predicting Breast Cancer Recurrence with Data Mining (2014)
- A Comparison of Selected Data Mining Algorithms (2014)
- Predicting Quantifiable Forest Fires using Data Mining
- Scooby-Doo: Where are You? : Examining the Dogs of New York City (2013)
- Drafting the Perfect Running Backs- Predicting Seasonal Performance Changes (2013)
- Predicting Meteorite Landings (2013)
- Predicting who Survived on the Titanic (2013)
- Micro-Loans: Predicting if Loan is Approved and Predicting Interest Rate of Loan (2013)
- Are there differences between religious and non-religious US Universities (2019)
- Predicting Shelter Animal Adoption: a look at characteristics that predict pet adoption (2019)
- Predicting AirBnb Prices (2019)
- Predicting Undergraduate Student GPA (2019)
- Predicting Success of Google Play Store Applications (2019)
- Determining Facial Emotions with Convolutional Neural Networks (2019)
- Interpreting Spam SMS Classifications using LIME (2019)
- Classifying Myoelectric Signals into Hand Movements for Control of a 3D Printed Hand (2019)
- Predicting E-sport Team Performance at the League of Legends World Championship (2019)
- IMDB Movie Review Sentiment Analysis (2019)