# CSC 535/635 – Data Mining Project Guide

The goal of the course project is to apply data mining techniques to real-world problems. It is recommended to use scikit-learn for the project. You can use any built-in IPython or Python functions in the project code. The use of other libraries may be allowed upon approval from the instructor.

The following guidelines will be used to evaluate projects:

* 1. Whether the motivation of the project is well described.
  2. Whether the dataset(s) is properly explained.
  3. Whether the choice of data mining techniques is well justified.
  4. How well the experimental results are discussed.
     1. Why does a method produce better results than others? (if applicable)
     2. Why do not the methods give significantly different performance? (if applicable)
  5. Whether the method is properly applied to realistic data set(s) that may be used in practice.
  6. How well the experimental settings are described.
     1. What preprocessing was done to the data? (if applicable)
     2. Was feature selection considered?
     3. What methods do you consider?
     4. How do you tune the optimal parameter(s)?
     5. What are the evaluation strategies (e.g., k-fold cross validation) for the performance comparison?
     6. How many times are the experiments repeated? (if applicable)
  7. How well the experimental results are discussed.
     1. How can the experimental results be interpreted?
     2. Are the results and data set properly visualized? (if applicable)
  8. Conclusion
  9. Misc.
     1. What obstacles have you overcome?
     2. Any issues that you want to share?
     3. Any future plans for related research?
  10. Code: Is the code easy to read and understand? Make sure to use good programming practices. Make sure that the code is well structured, cleaned (unnecessary blocks are removed), well documented and commented.

For the project proposal, find a dataset that you may be interested in. You can refer to homework 1 for possible places to look for datasets. Kaggle.com and UCI Machine Learning Repository contain a lot of interesting datasets. If you have any specific related dataset, you can suggest it too.

**The followings must be included in the proposal:**

1. Names of group members. **Each group will consist of three to four students**.
2. Problem definition
3. List of the method(s) that you will implement (or compare). You must use some data mining/ML algorithms in your project. Only doing data analysis using Pandas won’t be sufficient for the project but can be part of the project. This course does not teach deep learning, so do not plan on using deep learning for the course project. However, if you want to use Keras or TesnsorFlow for traditional NN, that is allowed. Do not plan on using Spark for the project.
4. Brief description of the data (e.g., # of samples and features) and meaning of attributes where needed. URL of the dataset if it is available online.
5. Description of data setting (e.g., Panda’s DataFrame, Excel file, …etc.) and preprocessing (if need)
6. It is assumed that students will use scikit-learn and IPython for the project. You must get prior approval from the instructor if you plan to use other libraries.

**Report Specifications:**

When writing the project report, please look at the above project guidelines so that you are aware how your project will be evaluated.

The length of the project report needs to be around 8 pages. In general, 6 to 10 pages will be fine. Marks will be deducted for exceeding the 10 pages limit. Please use 12pt Calibri font, 1-inch margins, **single column, double space** between the lines. Deviating from these guidelines will result in losing points.

Divide the report into sections as appropriate. For example, you can have sections for abstract, introduction, approach, experiment set up, results, conclusion, and references. You can add more sections as you feel appropriate. For groups that contain graduate students, I recommend following the format of a research paper. You can use Google to search for examples of research papers if you need but a sample is at https://irojournals.com/jscp/V2/I2/07.pdf.

In the introduction, introduce the problem. Give background information about the problem, if needed. Explain the dataset(s) used. Explain the meaning of attributes if needed. List the size of the dataset (number of samples and attributes). Talk about related work, if needed. This will include what approach others have used to solve this problem. Briefly talk about how your approach is different. Briefly compare your results to those of others, if needed.

In the approach section, write your approach for solving the problem and explain why you chose this approach. Give background information if needed. Explain any preprocessing you did to the dataset.

In the experiment section, you can write about hyperparameter tuning. Write down the tools that you used to write your code. It is recommended to show results in tables and/or graphs as appropriate. **Talk about the results and** **explain the results** if needed. Try to think why a certain algorithm performed better than another or why a certain algorithm, if any, did not give good results. Add a subsection with instructions on how to run your code. Seed random number generators where needed in your code so that we get the same results when running your code. Make sure that your code is commented and well documented.You can lose many points if your code is not commented.

In the conclusion section, you can draw conclusions about your project. You can write about what you would do if you had more time or if you redid the project from scratch. Write about what worked and what did not work.

It is a good idea to add an abstract before the introduction with a brief description of what you did for the project and the results. This should be one paragraph with no more than 10 lines.

In the references section, write the references to all sources other than the course material that you used. Cite these references in your report where needed.

**Add a section after conclusion and before references named Members Contributions**. List the contributions of each group member here. Generally, group members are expected to receive the same grade. However, if a member did not collaborate with other members or did not do enough work, then his/her grade will be different.

While writing the report, do not use scientific terms that were not explained in CSC 535/635 without explaining them. For example, if you write exploratory data analysis, you need to explain what this means. If you write linear model, you need to explain what this means. If you use algorithms/models that were not discussed in the course, you will need to explain in your report how these algorithms/models work to convince the instructor that you understand them.

**Presentation:**

You will create a presentation that is between 8 to 12 minutes in length. Points will be deducted if presentation is over 12 minutes. In the presentation, use PowerPoint slides to explain your project, the dataset(s) used, your approach, and discuss the results you got. The presentation **must include a demo of the code**. You must run the code as part of the presentation. You do not need to explain methods discussed in class. On the other hand, you will need to explain methods that were not discussed in class to convince the instructor that you understand them. It is preferrable if each group member talks during the presentation. For example, you can divide the presentation where each member presents his/her contribution. You can use any software you like to record the presentation. For example, you may use Zoom or Blackboard Collaborate Ultra.

**What to turn-in:**

Your project report, presentation, and code will be checked by Blackboard for originality. So, they will be uploaded as separate unzipped files in links created for that on Blackboard. Please upload the report as a Microsoft Document file. For the code, use a free online tool to save/convert as .pdf file before you upload it to Blackboard. The presentation video will also be uploaded to Blackboard. **If the video is too large, you can make it available online and provide a link in a separate file named video\_link**. Upload all files to Blackboard. Only one submission per group is needed. Please remember to include your names on all documents.