CSE 469 Project:
Solve a Data Mining Problem

Report and Presentation Due: December 5, 2023 11:59pm

In this project, you will practice what you are learning in class to solve a real-world data mining problem. You can choose any problem that you are interested in as long as it can be formulated as a data mining task, such as classification, association analysis, clustering, anomaly detection, or a combination of these. This project is a team project. Each team should have three members.

Complete the following tasks:

1. Pick a real-world application that data mining may help. It does not have to be ambitious.
2. Formulate it as a data mining problem (classification, association analysis, clustering, anomaly detection, or a combination of these tasks).
3. Collect relevant datasets. Some possible sources:
   - https://archive.ics.uci.edu/datasets
   - https://kdd.ics.uci.edu/
   - https://www.data.gov/
   - http://www.kdnuggets.com/datasets/index.html
   - https://www.kaggle.com/datasets
   - https://datasets-benchmarks-proceedings.neurips.cc/paper/2021 (novel datasets are provided in papers)

4. Preprocess the datasets into the format that can be used by data mining algorithms, if necessary.
5. Apply your implemented algorithms or use any existing package to solve the proposed problem. You can use any existing package to preprocess your data, implement the algorithm, postprocess your results, and prepare an appropriate visualization (if applicable). Some existing packages are (in no specific order or category) Scikit-learn, Matplotlib, Seaborn, NumPy, Pandas, SciPy, statsmodels, Weka, NLTK, GENSIM, Biopython, etc.
6. Discuss the data mining results you obtain and evaluate the results.
7. Prepare a short report based on the key points of your project. Name it as project.pdf. Include the names and UB emails of your team members at the top of your report. **Each team member should describe their contribution to the project.**

8. Submit your report to Autolab. Your report should include the following components.

   - Introduction: What data mining problem are you trying to solve? What impact will it bring if the problem is solved?
   - Formulation: Which data mining task it can be formulated into? What’s the input and the expected output?
   - Datasets: Where did you get the datasets? Give some statistics about the data. How did you preprocess the data?
   - Algorithm: Which data mining algorithm did you apply?
   - Experiments: Evaluate the output using an appropriate evaluation metric. Show the results you get and discuss whether they are meaningful.
   - (Optional) Challenges: What challenges do you find in the data? How can you tackle these challenges?

9. Each group will also make a short presentation in the class in the last week of the semester.