

Project: Solve a Real Data Mining Problem

Report Due: December 7 11:59pm

In this project, you will practice what you learn 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. This project is a team project. Each team should not have more than two members.

Complete the following tasks:

1. Pick a real-world application that data mining may help.
2. Formulate it as a data mining problem (clustering, classification, association analysis, anomaly detection, recommendation, or a combination of these tasks).
3. Collect relevant datasets. Some possible sources:
 - <https://archive.ics.uci.edu/ml/datasets.php>
 - <https://kdd.ics.uci.edu/>
 - <https://www.data.gov/>
 - <http://www.kdnuggets.com/datasets/index.html>
4. Preprocess the datasets into the format that can be used by data mining algorithms if necessary.
5. Apply your implemented algorithms or any existing package to solve the proposed problem.
6. Discuss the data mining results you obtain and evaluate the results.
7. Prepare for a short report based on the key points of your project. Name it as project.pdf or project.doc or project.docx
8. Log in any CSE department server and submit your report as follows:
submit_cse469 project.pdf

Your report should include the following components.

- Introduction: What data mining problem you are trying to solve? What impact it will 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 do you get the datasets? Give some statistics about the data. How do you preprocess the data?
- Algorithm: Which data mining algorithm do 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 do you tackle these challenges?