King Saud University
College of Computer and Information Sciences
Information Technology Department
IT326: Data Mining
Project

Project Description

The aim of this project is to apply the data mining techniques you learned in the class to some real-world dataset. You can choose any problem that you are interested in, and formalize it into a data mining task, find a dataset related to the problem, preprocess the data and then use R to apply the suitable data mining techniques to the chosen dataset. Also, you need to evaluate and compare the results of different data mining techniques and discuss your findings. Finally, you should submit a report, with your data and code and give a short presentation of your work.

1. **Form Groups** [Deadline: Feb 2, 2020]

A group of 4 students is required for the project. The names of the group members and the group leader must be registered in LMS by the deadline.

2. **Data Selection** [Deadline: Feb 9, 2020]

The following are sites that contains a list of datasets. You can select one of these datasets to work on, or can propose data of your own.

- UCI Machine Learning Repository http://archive.ics.uci.edu/ml/
- Kaggle: https://www.kaggle.com/datasets
- Kdnuggets: https://www.kdnuggets.com/datasets/index.html
- Data.World: https://data.world
- World Data Atlas: https://knoema.com/atlas

Your dataset has to be ***approved*** by your instructor.

3. **Data Summarization and Preprocessing [**Deadline: Mar 1, 2020]

Apply different plotting methods that would help you to understand your dataset, such as scatter, histogram and bar plot. Apply data preprocessing such as data cleaning or data transformation.

4. **Data Mining [**Deadline: Mar 22, 2020]

Apply data mining technique (classification or clustering) to your data, present and discuss the results.

5. **Final Report and Presentation** [Deadline: Apr 5,2020]

Submit the final report, data, and R code. For the presentation, you are expected to present your problem, your dataset, the data mining techniques you used, and briefly discuss your results and findings. The presentation will be 5-7 minutes long and all team members should participate.

Final Report Content

1. **Problem**

Introduce the problem. What do you want to solve? Why do you think it is important?

2. **Data Mining Task**

Formalize the problem as a data mining task

3. **Data**

Describe the dataset include: Source, Number of objects, Number of attributes, and the main characteristics of attributes (e.g. data types, distributions, missing values, etc..) using statistical measures and graphical presentation you learned.

4. **Data preprocessing**

Explain why did you (or didn't you) choose to apply data preprocessing. If your data required preprocessing, explain your process. Provide justification for the techniques you applied. In your submission, include your raw dataset, as well as your processed dataset.

5. **Data Mining Technique**

Describe the data mining technique that you will apply to your dataset and why.

6. **Evaluation and Comparison**

Present, evaluate and compare the result of different techniques applied to your dataset.

7. **Findings**

Discuss your findings by investigate the obtained mining results and decide whether these results are interested or not.

8. **Code**

Present the R code you write to accomplish the data mining task.

Good Luck!