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

1. Project Description

The aim of this project is to apply the data mining techniques you will learn 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 GitHub with jupyter notebooks discuss your finding on the dataset.

This project consists of 3 main phases, each with a specific task to be fulfilled. The deliverables of each phase will be added to **GitHub**. Thus, each group has to create a project in GitHub that is organized based on the deliverables of each phase. In this way, your instructor can easily find your submissions. The final phase will include giving a short presentation of your work and updating your project according to the instructor's feedback.

Hint:

- How to upload your R code on GitHub? An example with an R: https://statsandr.com/blog/how-to-upload-r-code-on-github-example-with-an-r-script-on-mac-os/
- How to create R notebook: https://rstudio-pubs-static.s3.amazonaws.com/256225_63ebef4029dd40ef8e3679f6cf200a5a.html

  - **Form Groups** [Deadline: Thursday 7-9-2023]

  A group of 3 to 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. Phase by Phase guide

1. **Data Selection** [Deadline: Thursday 14-9-2023]

In this phase, you should find a dataset to work on. You can select a dataset from the following sites, or you 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
- Google Dataset Search https://datasetsearch.research.google.com/

Please note that the dataset has to be approved by your instructor. So, you should choose an appropriate dataset that you understand, thus, you will be able to use it for a clear goal. Rows should be not less than 500 and columns not less than 10. Also, please check that there is no other group that has already chosen it.

Hint: make sure you can access and read the selected dataset in R languages.

Each group is required to submit GitHub project link that contains the following:

1. A suitable information about your project including students' names and project motivations (why you decide to work on this problem/dataset).

2. The dataset under a "Dataset" folder

1. R Jupyter notebook showing the description of the dataset, which should include the following:
   - The goal of collecting this dataset (e.g., iris classification, defect prediction …etc.)
   - The source of the dataset, include link\URL
   - General information about the dataset such as the number and the type of attributes, number of objects, the class name or the labels

2. Data Summarization and Preprocessing [Deadline: Monday 8-10-2023]

In this phase, you should apply different summarization and plotting methods to help you to understand your dataset, such as scatter, histogram and bar plot. Then, you will apply preprocess your data as needed using data cleaning, data transformation and feature selection.

Each group is required to submit GitHub project link that contains the following:

- A summary of the dataset including samples of raw dataset, graphs and tables show variables distribution, missing values and statistical summaries (Central tendency measurements such as mean and variance).

- The application of preprocessing techniques on your data with justifications: variable transformation, discretization, removing noise, feature selection, and normalization.

3. Data Mining [Deadline: Sunday 5-11-2023]

In this phase, you should apply data mining techniques (classification and clustering) to your data, present and discuss the results.

1. Classification:

Apply Decision trees classification to your dataset using at least three different size of partitions and three attribute selection measures (IG, IG ratio, Gini index). Explain your reasons to select certain partition method then, compare and discuss the results on your dataset. Use visual representation and try your best to interpret the results and to understand the algorithms performance.

2. Clustering:

Apply K-means clustering to your dataset using at least three different sizes of K. You have to justify your choice for size K then compare and discuss the results using different evaluation methods and metrics

(Silhouette coefficient, total within-cluster sum of square, BCubed precision and recall). Use visual representation and try your best to interpret the results and to understand the algorithms performance.

4. **Final Report and Presentation** [Deadline: Sunday 26-11-2023]

Finally, you should submit a complete report by converting your GitHub notebook to pdf, and present your work to your classmates. 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 Contents

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 techniques (classification and clustering) 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!