THE UNIVERSITY OF HONG KONG

DEPARTMENT OF STATISTICS AND ACTUARIAL SCIENCE

STAT8017 Data Mining Techniques (2nd Semester 2020-21)

Group Project

Purpose:

This project aims to provide students with more practical experience of using data mining tools on a real-life problem. You will formulate a problem and apply relevant data mining tools in practice. The project will account for 30% of total assessment.

Project Teams:

Each team consists of 2 to 4 students. You form a team yourselves and email me before the 1st class test, including project title and team members.

Details of Projects:

1. Identify a topic and determine the objectives. You can use one of the following Kaggle datasets:

https://www.kaggle.com/ksaivenketpatro/fake-news-detection-dataset

https://www.kaggle.com/qbatista/us-stocks

https://www.kaggle.com/burmad/patient

https://www.kaggle.com/econdata/predciting-price-transaction

https://www.kaggle.com/isadoraamorim/traffic-crashes-crashes

https://www.kaggle.com/frabbisw/winedata

https://www.kaggle.com/m0hd7ah1r/bollywood-movie-dataset

https://www.kaggle.com/man0007/churn-modelling

https://www.kaggle.com/sulianova/cardiovascular-disease-dataset

https://www.kaggle.com/abdurrehmankhalid/delayedflights

https://www.kaggle.com/sherinclaudia/nyc311-2010

https://www.kaggle.com/miteshsingh/hollywood-music-dataset#Hollywood-Music-WCBS-Ranking.csv

https://www.kaggle.com/acmfootball/premier-league-football-data-200708201718

https://www.kaggle.com/aashishmalik7936/data-driven-comp#train_values.zip

https://www.kaggle.com/masaladata/14-million-cell-phone-reviews#phone_user_review_file_2.csv

https://www.kaggle.com/jjingmsba/restaurant-score

https://www.kaggle.com/budhajit/plane-crash-information-dataset

https://www.kaggle.com/bobirino/data-file-feature-engineering-tutorial

https://www.kaggle.com/sparnord/danish-atm-transactions

https://www.kaggle.com/kuncoroaji/weather-aus

https://www.kaggle.com/lpdataninja/dj-trump-tweets

https://www.kaggle.com/tylerx/melbourne-airbnb-open-data#reviews.csv

Alternatively, you can subscribe to be a Kaggle member and access hundreds of datasets in their machine learning data repository. Of course, you can always use your own data source (e.g., workplace data subject to corporate restrictions)

2. Explore the dataset. Pay attention to the quality of data (e.g. missing values), the meaningful features, data distribution, and types of variable. Perform data cleansing and transformation.

3. Choose appropriate DM techniques and develop a DM model over the dataset. You can also use a new DM technique outside the syllabus as long as you clearly describe the methodology.

4. Fine-tune the model and explain the outcomes with regard to the project objectives.

5. Each team must submit a project report (in pdf), including python code (in both .ipynb and .html) and the dataset (eg, txt or csv, etc) via the moodle submission area (will be set up later).

Project report:

The report can be submitted any time before the exam period starts. This year, S2 exam period starts from May 10, 2021. Hence, you can submit up to ***11:59pm on Sunday, May 9, 2021***. You report should include the title, abstract, motivation, model, data, methodology, analysis, discussion, conclusion, references (eg, websites, book chapters, articles, etc) and the computer code/output. A table of contents should also be given at the start of report. A subject index (giving the page numbers for key words used) at the end of the report is also preferred. The report has no upper or lower limit on word length, as long as the requirements above are satisfied. The following are useful tips to keep in mind:

- Project objectives (background, the problem, and purpose).
- Data description/preprocessing (source of data, description of major features/variables, quality of the data, and appropriate data preparation).
- DM results (the results, and other DM techniques to be considered, describing the problem to be encountered and how they might be solved, etc.)