STAT3612 Statistical Machine Learning
2021/2022 Semester I
Group Project

Overview

The project aims to provide you with more practical experience of using machine learning tools and R skills learned from the class on a real-life problem. You will learn how to formulate a problem, apply relevant machine learning tools in practice and write a report. Each project should be done in a team of 3 to 4 members. The report will account for 30% of your final grade.

Details

1. Identify a project topic and determine the objectives of the project. Find an available data set. Some sources of data sets:
   - Hong Kong open data: https://data.gov.hk/en/
   - Kaggle datasets: https://www.kaggle.com/datasets
   - UCI machine learning repository: https://archive.ics.uci.edu/ml/index.php
2. Study and understand the data set. Pay attention to the quality of data (eg. any missing values), the meaningful attributes (variables), data distribution, and the types of variable values. Perform necessary cleansing and transformation.
3. Choose appropriate machine learning techniques and use R to develop a machine learning model on the data set.
4. Fine tune the models and interpret the results as much as possible with regard to the project objectives.

Report

1. The report should be 7-10 pages long (excluding references and appendix) and contain:
   - objectives of the project (including the background of the project, the project problem, and project objectives)
   - description of data and data preparation (including the source of data, the description of major attributes (variables), the quality of the data, and data cleansing and/or transformation)

- description of machine learning techniques used and how the techniques solve the project problem
- interpretation the results of the analysis
- conclusions and limitations of project
- references and appendix (including codes, tables and figures)

2. Grading is based on:
   - creativity and problem difficulty (10%)
   - correctness of analysis and interpretation of results (10%)
   - organization and clarity of writing (10%)
3. The report should be submitted through Moodle by 18:00 December 7, 2021. Late submission will not be accepted.