Instructions to Group/Individual Project

Due date: Saturday, August 10, 2024

Instruction:

- Project submission deadline: Saturday, August 10, 2024, by 11:59 pm central time
- Poster submission deadline: Wednesday, August 7, 2024, by 11:59 pm central time
- Discussions: You are required to read your fellow groups' posters and write your comments/questions/suggestions in the Discussion Board. The posters will be available from Wednesday, August 7, till Saturday, August 10, 2024.
- Project grading points breakdown: total 50 points
  - Written Project: 30
  - Poster: 15
  - Discussions: 5

In this project, you should team up with students who have the same intention. Each team can have 1–3 students (no more than 3). Each individual/group needs to various predictive modeling methods that you learned in this class to predict the onset of diabetes in Pima Indians given medical details. By the deadline of the project, you need to produce three documents:

- A written project in the pdf format, which will be submitted via the Project folder on Canvas.
- A complete R code that could replicate your results in your report, which will also be submitted via the Project folder on Canvas.
- A poster in the pdf format, which will be submitted to the Poster Submission and Discussion folder on Canvas.

Each team (from one to three members) is required to analyze one of the following two data sets:

- Data set 1: The Pima Indians Diabetes Database can be found here.
- Data set 2: The House Prices Data can be found here.

Requirements of the Data and Analysis:

All the methods that you are using should be the ones that you learned or will learn from this course. Keep in mind that data pre-processing may be needed before applying statistical machine learning methods if necessary. Several methods/models that you may consider for analyzing this data include but not limited to the following:

- Linear regression and its cousins
- Multivariate adaptive regression splines
- Logistic regression
- Discriminant analysis (linear and nonlinear)
- Decision trees
- Flexible discriminant analysis

- Penalized models and Nearest Shrunken Centroids
- K-nearest neighbors
- Bagging and random forest
- Boosting
- Support vector machine
- Neural network
- Other models.

The Objective of the Project and Your Role:

The objective of this project is for students to do data analysis by applying various predictive modeling techniques learned in the class. You can choose any of these appropriate methods based on your interests and knowledge. (Hint: different data types may require different models to analyze.)

View yourself as a consultant to your client who presented you this data set. When you run the analysis and work on the write-up, always think about how you would convey your idea and results to your client. It may be a good idea not to assume your client knows much of statistics and/or predictive modeling methods. As always, I will be available to discuss your projects or to answer your questions, if any, as well.

This is an individua/group project so within each group the members should communicate with each other and contribute as much as you can. In addition, I will be available for discussing your projects or answering questions as well.

The Recommended Layout of the Project:

(1) Introduction and Background: describe the nature, goal, and brief description of your analysis methods clearly and sound; Objectives should be clearly stated

(2) Data Structure: provide a clear picture of how the measurements are made and how the data are collected

(3) Statistical Learning Method(s): discuss concisely how the method is good in the study and why the the method can be used to reach your goal

(4) Analysis Results: display your result clearly and readable

(5) Conclusions, Discussions, and Bibliography: clearly state your conclusions in plain English. Whenever it is necessary, cite your references.

Other Project Requirements:

(1) A logical and well-written analysis and solution of the problem

(2) A well-commented computer code performing all of the computations and producing all the output and graphics necessary for the paper (or your output, please edit it in a fashion like those appeared in a published paper. Do not just "copy and paste.")

(3) The total length of your project should not exceed 10 pages (excluding graphs and codes)

(4) Preferred using LeTeX to write your report, but the one from word documentation is also fine.

(5) Do not include any R code in your report. Submit your final project report and your R code with two separate files.

Keep in Mind:

(1) The Canvas plagiarism tools will be used to check your writing so do not copy and paste.

(2) Late submissions (penalty or not) are not accepted after the hard deadlines mentioned above.