Stat 427/627, Statistical Machine Learning, Course Projects **Summer 2024**
Due date: Pending. No later than Friday, June 28, 2024

You will work in groups of 2 – 4 on a course project. In this project, you will:
- Apply sound statistical machine learning techniques to a real-world data set.
- Tune the model for better accuracy.
- Assess the model(s) for performance.
- Summarize your findings and make recommendations.
- Provide reproducible code to support your conclusion.

Data Source
- You will choose a real-world data set from the following or other approved source (do not use Kaggle).
  - UCI Machine Learning Repository, https://archive.ics.uci.edu/ml/index.php
  - Government data: https://www.data.gov/ (enter a key search phrase, such as "racial profiling"):
  - US Bureau of Justice Statistics (suggested by our recent colloquium speaker): https://www.bjs.gov/index.cfm?ty=dca
- The data set should:
  - Have at least 500 sample size (after removing missing values)
  - Have at least 8 variables of interest.
  - Take caution with time series data.
- If you choose your data set outside the above sources, the data set must not contain any identifiable personal information, classified or proprietary data.
- Show me your data set and a project plan before you start the analysis.

Data preparation and exploration
- Remove all *missing values*.
- Randomly split your data set to set aside 20% of the observations as the *hold-out testing* data. These observations will NOT be used for model fitting and tuning. We will treat them as if they become available after you have fitted and tuned your model.
- Conduct an *exploratory data analysis*.

Fit and tune the model
- Use at *least FOUR different statistical learning algorithms* covered in the course across the regression and/or classification questions. You can choose from linear, logistic, polynomial regression with proper variable selection, linear or quadratic discriminant analysis, K-nearest neighbor classifier, jackknife, bootstrap, ridge regression, lasso, principal components regression, partial least squares, splines, regression and classification trees, artificial neural networks, support vector machines, clustering, or related methods.
- Apply *cross-validation* techniques to tune the model, such as the optimal degree of flexibility, the best subset of predictors, or the optimal tuning parameters.
- *Illustrate* results with appropriate tables, plots, and diagrams.
- *Evaluate* performance of competing methods. Also comment on the advantages and restrictions of the methods. Then make *recommendations*.

Test on the hold-out testing data

- Be sure that your estimated and tuned models have not "seen" the hold-out testing data before this step.
- Applied the competing models to the hold-out testing data. Then evaluate the prediction accuracy. Do the models perform as expected?

Presentation

Prepare and give a 10-15 minute presentation of your work. Your presentation should include, at minimum,

1. A brief background of the data set.
2. Exploratory data analysis.
3. Implementing the machine learning algorithms
    a. Fitting the models.
    b. Tuning the models: what to be tuned and how to tune.
    c. Model performance assessment.
4. Summary of your findings and recommendations.
5. Testing the fitted and tuned models to the hold-out testing data. If they do not perform as expected, try to explain the (seemingly) contradictory.

Teamwork

You may divide the workload within your team. Each team member is expected to contribute to the project. In addition, each team member is expected to be familiar with every aspect of the project even if he/she does not work on that portion. A within group self-evaluation will be conducted and accounted towards the project grade if needed.

Expected progress

Form your group(s) and discuss your project data.	Friday, May 31.
Exploratory data analysis, decide the variables.	Friday, June 7
Project outline, start fitting models.	Friday, June 14
Continue model fitting. Draft slides done.	Friday, June 21
Presentation (in-class or by recording TBD)	Tuesday or Thursday, June 25, 27.

Deliverables

a. The original data set that you analyzed.
b. R code (R script or R markdown) that replicates your analysis in full and includes all data processing, decision points, and your analysis. Include brief comments of main steps to make the code readable. This file should run or compile without error and include a statement or statements to load your data. You should set a random seed (e.g., set.seed() function) any time you randomize, such as splitting the data set so that the results are reproducible.
c. Your presentation slides, or in another format if you choose to.

(The assignment instructions end here.)

Other Data Sources (Just for your reference. Do NOT spend too much time searching for data set this course project. Summer schedule is tight.)

There are many sources you can find many relatively large and complicated data sets. Here are a few examples.

- COVID-19 data repositories such as:
  - https://github.com/nytimes/covid-19-data
  - https://www.cebm.net/oxford-covid-19-evidence-service/
- Race and Economic Opportunity data tables (US Census)
  - https://www.census.gov/programs-surveys/ces/data/public-use-data/race-and-economic-opportunity-data-tables.html
- Employment data by occupation, sex, race, ethnicity (US Bureau of Labor Statistics)
  - https://www.bls.gov/cps/cpsaat11.htm
- US Government's open data (enter a key search phrase, such as "racial profiling"):
  - https://www.data.gov/
- Open data portals from state agencies such as:
  - https://opendata.dc.gov/
  - https://ohio.gov/wps/portal/gov/site/government/resources/ohio-data-analytics
- Federal bridge inspections from the Federal Highway Admin:
  - https://www.fhwa.dot.gov/bridge/nbi/ascii.cfm
- Drinking water violations from the EPA:
  - https://www.epa.gov/ground-water-and-drinking-water/drinking-water-data-and-reports
- Contracts or grants from USA Spending:
  - https://www.usaspending.gov/#/
- DC Crime incident data from the DC police:
  - https://mpdc.dc.gov/page/statistics-and-data
- ~~SBA disaster loans from the Small Business Administration:~~
  - ~~https://www.sba.gov/offices/headquarters/oda/resources/1407821~~
- Fatal accidents from the NHTSA:
  - https://www.nhtsa.gov/research-data/fatality-analysis-reporting-system-fars
- Baltimore crime data from the Baltimore police:
  - https://www.baltimorepolice.org/crime-stats/open-data
- Aircraft wildlife strikes from the FAA:
  - https://wildlife.faa.gov/databaseSearch.aspx

- Bank health data from the FDIC:
  - https://www.fdic.gov/bank/statistical/guide/data.html
- USGS water quality data:
  - https://water.usgs.gov/owq/data.html
- Prison data:
  - http://www.dc.state.fl.us/pub/obis_request.html
- College Scorecard:
  - https://catalog.data.gov/dataset/college-scorecard-c25e9