

Final Course Project - Due Date April 30, 2021

Note: Students may work on this final project in groups of **at most 4 students**. Your submitted project **must contain statements outlining who was responsible for which part of the project**.

Instructions:

1. See links below for possible data sets to analyse; choose only **ONE**.
2. Your project **must** include application of some pertinent methodologies from **each** of the following topics:
 - visualization
 - dimension reduction
 - data reduction
 - unsupervised learning
 - supervised learning
3. Your project must be presented in written form, with clear statements about the methods used and your findings.
4. Code must be provided as an appendix to your report and in a format I can run to verify your results.
5. Your projects are to be submitted to me electronically on CuLearn.
6. **No late projects will be graded.**

Grading:

Each project will be graded (generally with all group participants receiving the same grade) taking into account the following criteria:

- Substance: Did you provide interesting, useful information that shows the "big picture?"
- Clarity: Did you explain clearly and concisely the main issues, and what your conclusions are?
- Focus: Did you cover the main points of the topic fully and without superfluous discussion?
- Presentation: Was the presentation professional?
- Completeness: You must provide the code used to obtain your analyses, as well as a summary write-up of your results.

Slackers: If you decide to work in a group, try to resolve workload issues within your group, but if it is clear that an individual is not pulling their weight I reserve the right to reduce his/her resulting grade.

Datasets:

Mushrooms Dataset It contains these columns: class, cap-shape, cap-surface, cap-color, bruises, odor, gill-attachment, gill-spacing, gill-size, gill-color, stalk-shape, stalk-root, stalk-surface-above-ring, stalk-surface-below-ring, stalk-color-above-ring, stalk-color-below-ring, veil-type, veil-color, ring-number, ring-type, spore-print-color, population, habitat. Here is [the link to this dataset](#)

NHANES Dataset The column names of this dataset may not look very understandable at first. [Download it from here](#) .

Housing Price dataset – commonly used to predict price This dataset contains these columns: id, date, price, bedrooms, bathrooms, sqft_living, sqft_lot, floors, waterfront, view, condition, grade, sqft_above, sqft_basement, yr_built, yr_renovated, zip code, lat, long, sqft_living15, sqft_lot15. Here is the [link](#).

Heart Disease (Cleveland) <http://archive.ics.uci.edu/ml/datasets/Heart+Disease>

Hypothyroid <http://archive.ics.uci.edu/ml/datasets/Thyroid+Disease>

Census income <http://archive.ics.uci.edu/ml/datasets/Census+Income>

Credit approval <http://archive.ics.uci.edu/ml/datasets/Credit+Approval>