CPSC 436/536 Final project

Project Description:

The goal of the final project is to integrate and apply the machine learning techniques for classification/regression covered throughout the semester to a real-world dataset of your choice. You will demonstrate your ability to perform data preprocessing, model selection, performance evaluation, and interpret model outputs. Make sure your dataset is interesting and big enough to showcase the ML techniques you've learned.

For this project, you will work in a group of 2-3 students. You can divide work by strengths, but each one of you must contribute to the technical work and the final presentation as well as an **individual report**. Each submission must include a short paragraph describing **each member's contribution**.

Project Components:

- Problem Description and Data Exploration
  1. Clearly define your prediction or discovery problem.
  2. Provide dataset description (source, variables, and size).
  3. Conduct exploratory data analysis: summary statistics, visualizations, feature relationships.
  4. Discuss potential challenges, missing data, or biases.

- Preprocessing
  1. Handle missing values, outliers, and categorical features
  2. Scale or normalize numerical variables
  3. Create or select features when relevant

- Model Selection and Training
  1. Train at least three ML models before you finalize your selection
  2. Use cross-validation and hyperparameter tuning (e.g., GridSearchCV)

- Evaluation
  1. Classification metrics: accuracy, precision, recall, F1, ROC-AUC, confusion matrix (heatmap)
  2. Regression metrics: MAE, RMSE, R^2
  3. Utilize PCA for your visualization

- Consider using unsupervised Learning to see if there are unlabeled classes
  1. Clustering (k-means, hierarchical)
  2. Discuss how this complements your supervised results

- **Discussion & Interpretation**
  1. Which model did you select, and why?
  2. What features were most important or influential? Do they make sense?
  3. How do model stability/overfitting appear in your results?
  4. What would you do differently with more time or data?

Datasets:

1) You own lab datasets
2) Health: NHANES, diabetes, heart disease, breast cancer
3) Finance: credit approval, loan default, bank marketing
4) Image/text: handwritten digits, IMDB reviews, news articles

Some Dataset Repositories:

- UCI Machine Learning Repository: Classic ML datasets with clear documentation and varied domains. https://archive.ics.uci.edu/ml/index.php
- Kaggle Datasets: Large community-driven dataset hub with ready-to-use CSVs and notebooks. https://www.kaggle.com/datasets
    - Insurance Claims Dataset (Kaggle): Structured dataset for predicting whether an insurance claim will occur. Ideal for binary classification practice. https://www.kaggle.com/datasets/litvinenko630/insurance-claims
- Google Dataset Search: Search engine for public datasets across the web. https://datasetsearch.research.google.com/
- Data.gov: Official U.S. open data portal with thousands of government datasets. https://www.data.gov/
- OpenML: Collaborative online platform for ML datasets and benchmarks. https://www.openml.org/
- KDnuggets Dataset Directory: Curated list of links for public datasets in many fields. https://www.kdnuggets.com/datasets/index.html

Dataset Selection Tips:

1. Avoid overly small datasets (e.g., < 100 rows), would be hard to evaluate multiple models.
2. Avoid massive raw datasets unless you have time to extract features and preprocess them.
3. Choose data that's tabular (rows × features) unless you're ready for text/image data.
4. Make sure your dataset has clear labels if you work on supervised ML.
5. Always include a citation and dataset link in your report.