CPSC436/536 Project2
Decision Trees, SVMs and Ensemble Learning

Objective: Understanding the concepts of decision trees, SVMs, random forest, gradient boosting, and how to use them in real world applications.

Dataset: The same data set, except the target is **stroke**. You'll work on a dataset (attached) extracted from National Health and Nutrition Examination Survey: https://www.cdc.gov/nchs/nhanes/index.htm.

- The dataset contained health records of n NHANES participants. The attributes include:
    - Age, Gender, Race, Blood Pressure readings (Systolic & Diastolic), Lab work (levels of total cholesterol (TCHOL), LDL, HDL, triglyceride), and certain medical conditions such as Diabetes. We also know whether he/she is a current smoker (smoker).
    - In addition to the above attributes, medical professionals consider some cross terms are important, such as age* Systolic, age* TCHOL, age*HDL, age* smoker. You might want to consider them.
    - Target variable: **Stroke**. (whether the participant had a stroke).

Goal: Predict the **probability** of a participant who had suffered a stroke.

Output: Utilize your <u>most optimal models</u> to forecast the likelihood of individuals in the testing dataset who had suffered a stroke.

Evaluation: Your project will be graded based on the difference between your predicted probability and the true label. Specifically, I'll be using both the accuracy and my modified Kullback-Leibler (KL) divergence between the predicted probability and the observed target,

$$D_{MKL}(P,Q) = \left| \frac{1}{2} \sum_{x \in X} (P(x) \log \left(\frac{P(x)}{Q(x)} \right) + Q(x) \log \left(\frac{Q(x)}{P(x)} \right)) \right|$$

where P(x) is the true probability, Q(x) is your predicted probability.

Things to consider when you tune your models:

1. How many features/attributes does the dataset have? What is the class distribution?
2. How many instances are in class1 and how many in class2? If it's unbalanced, should you consider balancing the data? There are some Nas, how would you handle missing values?
3. Do you need to z-transform your dataset?
4. What are the best values for the hyper parameters, such as Gini impurity vs entropy, number of trees in random forest/gradient boosting, regularization parameter λ, kernel type, …?

What to submit: Your Jupyter notebook and your predictions for the participants in the testing dataset.

1. Project2.jpynb
2. randomforest_pred.csv
3. gradient boosting_pred.csv
4. SVM_pred.csv

(**Graduate students**) A report that summaries your investigation, and your understanding why some models perform better than others (2 pages)

Attributes keys:

Age	Continues
BMI	Continues
CurrentSmoker	1 yes; 2 no
Diabetes	1 yes; 2 no
Diastolic	Continues
Edu	1- Less than 9th grade; 2- 9-11th grade (Includes 12th grade with no diploma); 3- High school graduate/GED or equivalent; 4- Some college or AA degree; 5- College graduate or above
HDL	Continues
Income	Ratio of family income to poverty
isActive	1 yes; 2 no
isInsured	1 yes; 2 no
kidneys_eGFR	Continues
LDL	Continues
Pulse	Continues
Race*	1 Mexican American, 2 Other Hispanic, 3 Non-Hispanic White, 4 Non-Hispanic Black, 5. None, 6. Non-Hispanic Asian, 7. Other Race - Including Multi-Racial
Sex	1 male; 2 female
Systolic	Continues
TCHOL	Continues
Trig	Continues

*Sometime people consider only three race groups: white, black, and others