Fall 2025 CPSC436/536 Project2 - Logistic Regression, Random Forest, and SVM

Objective: Gain understanding of logistic regression, random forest, and support vector machines models in general and build, tune, and compare the three supervised classifiers to predict the probability that an NHANES participant has experienced a stroke.

Dataset: 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 attribute list includes:
        - age, gender, race, blood pressure readings (systolic and 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 interaction 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 has had a stroke).

Goal: Predict the probability of a participant who experienced a stroke, p(stroke=1|X).

Output: Utilize your most optimal model for each of the 3 ML models to forecast the likelihood of individuals in the testing dataset who had experienced a stroke.

Things to Consider:

1. Go over what you did for project 1 and see if you need to repeat some steps, such as feature selection, scaling, handling missing values, balance dataset, categorical variables encoding, feature engineering, … …
2. Tuning Logistic Regression (baseline + regularization)
    - Regularization search: Tune C (e.g., [0.001, 0.01, 0.1, 1, 10]) and penalty (l2, …).
    - …
    - Inspect feature importance
3. Tuning Random Forest
    - n_estimators (e.g., 100–500),
    - max_depth (None or 4–30),
    - min_samples_leaf (1–10),
    - …
    - Inspect feature importance.
4. Tuning SVM
    - Different kernels (linear, polynomial, RBF, …),
    - Regularization parameter C (e.g., [0.1, 1, 10, 100]),
    - Kernel coefficient gamma (e.g., ['scale', 0.01, 0.1, 1]),
    - …
    - Inspect feature importance

What to submit: Your Jupyter notebook and a .csv file with 4 columns of predicted probabilities for stroke = 1 for the participants in the testing dataset. For example,

Participant ID	Logistic regression prediction p(stroke=1|X)	Random forest prediction p(stroke=1|X)	SVM prediction p(stroke=1|X)
101	0.61	0.45	0.70
…			

Evaluation: Your project will be evaluated based on how well your model predicts the probability of a stroke. Specifically, we will use two metrics:

- **Accuracy**: Measures how often the model's predicted class (based on a 0.5 threshold) matches the true label.
- **Kullback-Leibler (KL) Divergence**: Measures how close your predicted **probability distribution** is to the true label distribution. Lower values indicate better calibrated probabilistic predictions.

$$D_{\mathrm{KL}}(P \parallel Q) = \sum_{x \in \mathcal{X}} P(x) \log\left(\frac{P(x)}{Q(x)}\right),$$

where $P(x)$ is the true distribution, $Q(x)$ is the predicted distribution. This is equivalent to log loss (cross-entropy), where $\hat{p} = Q(x)$ is the predicted probability:

$$D_{KL}(P \parallel Q) = y \cdot log\left(\frac{1}{\hat{p}}\right) + (1 - y) \cdot log\left(\frac{1}{1 - \hat{p}}\right)$$

You can try this metric yourself:

```
from sklearn.metrics import log_loss
kl_div = log_loss(y_true, y_pred_prob)
```

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
Insurance	Categorical
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. Blank for missing values.