**Data set description:**

There are two data sets named “origin\_breast\_cancer\_data.csv” and “breast\_cancer\_data\_357B\_100M.csv”.

* The data set “origin\_breast\_cancer\_data.csv” contains 357 benign and 212 malignant samples.
* The data set “breast\_cancer\_data\_357B\_100M.csv” contains 357 benign and 100 malignant samples, which is unbalanced for positive and negative samples.

Both data sets have 32 columns, starting from 1) to 23), which are explained below.

1) ID number (You can ignore it)

2) Diagnosis (M = malignant, B = benign)

3-32) Attribute Information: Ten real-valued features are computed for each cell nucleus.

a) radius (mean of distances from center to points on the perimeter)

b) texture (standard deviation of gray-scale values)

c) perimeter

d) area

e) smoothness (local variation in radius lengths)

f) compactness (perimeter^2 / area - 1.0)

g) concavity (severity of concave portions of the contour)

h) concave points (number of concave portions of the contour)

i) symmetry

j) fractal dimension ("coastline approximation" - 1)

The mean, standard error and "worst" or largest (mean of the three largest values) of these features were computed for each image, resulting in 30 features. For instance, field 3 is Mean Radius, field 13 is Radius SE, field 23 is Worst Radius. All feature values are recorded with four significant digits.

**Project description:**

1. Use the logistic regression method and kNN method for the binary classification problem (to predict whether it is malignant or benign) with the data set “origin\_breast\_cancer\_data.csv”. Try your best to pre-process the data and tune possible hyper-parameters, to get your best binary classification results for each method. Show the metrics (recall, precision, and F1 score) for both training set and validation set. Compare these two methods and discuss for the conclusions that you have.
2. Considering the unbalanced sample data set “breast\_cancer\_data\_357B\_100M.csv”, try the two methods with a new training process again to see what will happen. If the performance is degraded, then try some modification (your input and your creative ideas) for these two methods to see if you can have some improvement. Detail your ideas and show the performance you obtained.

**Requirement:**

1. You should complete your project by yourself only, and then hand in your code as well as a project report before the deadline.
2. The midterm project takes 20 marks for the course (20%).
3. The deadline to submit your project report and code packages is 23:59PM of Dec. 04, 2022. It is a firm deadline (Late submission will receive 0 mark).
4. When completing your course project, you are required to write a project report together with the codes for the project. Base on the project report and the code package, the project will be marked.
5. The project report should be written in English.
6. The project report should be presented in the IEEE conference paper style and suggest to use LaTex if possible. Refer to the following link

https://www.ieee.org/conferences/publishing/templates.html

for the LaTex Template (a LaTex template package is also included in the zipped file), or you can work in Overleaf (an online LaTex editor). The project report should contain the project title, authors, abstract, keywords, I. Introduction, II. Problem formulation, III. Method and algorithms, IV. Experiment results and analysis, V. Conclusion and future problems, and References.

1. Hand in a complete code package including the data set, code files with detailed description of dependencies, etc., so that the code can be checked and run on another computer without any problem.
2. The project and the codes should not be copied from others. Once it is noticed that the hand-in is copied from others including your classmates or online available work, you will receive 0 mark.

1. Mark criteria
   1. Creativity (5 marks): You have to have your own idea and input to solve the problem and analyze the results. Please be aware that the only way TA can understand your new inputs is from the project report. So please make sure that you have provided sufficient evidence to show your effort in the project report.
   2. Completeness (5 marks): The project should be a self-contained one. It should have a clear problem formulation, followed by a complete solution and algorithm, as well as experiment results and analysis, and concluding discussion.
   3. Presentation of project report (5 marks): The report should be well organized and clearly written. The problem under consideration and the developed solution (algorithm) should be clearly described. The simulation (experiment) results should be fully discussed and analyzed. Valuable conclusion should be provided. Any unclear points will get some marks off.
   4. Presentation of the codes and codes comments (5 marks): The codes should meet a good Python coding style and easy for reading. Also, the codes should be accompanied with clear and detailed code comments. Any confusion in understanding the codes may lead to some marks off.