Project 5 (Final Project): Binary Classification to Predict the Presence or Absence of Breast Cancer

Due: 11:59 pm, Dec 11, 2021

For Project 5 your assignment is to analyze data from the Wisconsin Breast Cancer Dataset:

- For graduate students (CPSC 6430), you are **required** to use k Nearest Neighbor, Logistic Regression, Support Vector Machine, and Multilayer Perceptron (**4** different algorithms)
- For undergraduate students (CPSC 4430), you are **required** to use k Nearest Neighbor, Logistic Regression, and Support Vector Machine (**3** different algorithms). If you include an additional solution using *Multilayer Perceptron (MLP),* you will get an extra ***5 points*** to your *final grades* of this course (your final grades of this semester + 5 points).

Note: You can use **Scikit-learn library** to implement all abovementioned machine learning algorithms.

The python command to obtain the data is:

```
from sklearn.datasets import load_breast_cancer
```

For each of the four ML algorithms you should print out a confusion matrix, accuracy, precision, recall and f1 metrics. Use a 75%/25% split between training and testing files. Use 0 as the seed for the randomization of your data.

What to turn in:

1. One py program that gets the data and computes a confusion matrix and metrics for all three ML algorithms (yourlastname_yourfirstname_P5.py)
2. A pdf file containing ((yourlastname_yourfirstname_P5.pdf)
    a. Your name and a title for your write up
    b. Confusion matrix for each ML algorithm
    c. Accuracy, precision, recall and f1 metrics for each ML algorithm
    d. Number of neighbors you used for kNN
    e. Number of iterations you used for Logistic Regression, SVM, and MLP (if you use)
    f. A one paragraph justification for which ML algorithm you think is the best choice for breast cancer detection.

Zip the two files before uploading.

Notes:

- Be sure you understand whether 0 or 1 indicates the presence of cancer and that you label your confusion matrix correctly.
- Be careful that you interpret precision and recall produced by the library correctly. What do they mean by TP, presence, or absence of cancer?