Final project – Logistic Regression, SVM and Neural Networks

For project 1, you selected a data set and investigated how kNN classifier could help with classifying testing samples after training/learning. For project 2, you will be continuing your exploration with other supervised ML approaches – **logistic regression, support vector machine, and Neural Nets**.

This project is to be done in a group. Each student is responsible for contributing to the group, including problem formulation, dataset selection, ML tool implementation, and project presentation.

Requirement:

1. Source code in Python (done in a group). Your code must run on CS lab machines.
2. **Individual project report** (~6 pages + appendices if needed, fonts>=11)

Specific requirements:

Dataset:

1. Each group needs to pick a **new dataset** to work on.
2. Dataset must be interesting and **challenging** (if the accuracy is very high, say 99% using a knn or very low (<50%), select a different dataset! That means either the problem can be solved without any machine learning algorithm or beyond what we have learned in this class.)

Your individual report that includes:

Abstract - Give a brief presentation of the **problem**, dataset used, summarize the methods, and outline your results and conclusions.

1. **Introduction** - Detailed problem description and background of the dataset. Justify the dataset is appropriate and worth to explore. Outline approaches you take to solve the problem.
2. **Statistical summary of your data** - For **each class**, what are: max, min, mean, median, mode, standard deviation. If you used only a subset of attributes, justify why other attributes were not used. Summary what the statistics tells you, any insights you have obtained from the statistics.
3. **Methods** - A brief description of each model, logistic regression, support vector machine (linear kernel), and neural networks. Also include what ranges of parameters and neural network architectures (consider at least 2 different hidden layers with different # of neurons and 2 different gradient decent solvers) you'll consider exploring and why? Demonstrate you have an intuitive understanding of the ML algorithms.
4. **Results** - Summary of your classification results, including **best set of parameters and architectures**, accuracy, and confusion matrices from a) logistic regression, b) SVM, and c) neural nets.
5. **Discussion** - Describe and analyze the results. Are the results what you expected? How do the three different models compare? Why one is better or worse than another?

6. **Conclusion** of your exploration. Did you solve the problem? How helpful are the ML algorithms in terms of answering your questions? What have you learned?
7. (**Graduate student**) Give a more detailed description of each model, logistic regression, support vector machine, and neural nets, and compare SVM with linear kernel and SVM with Gaussian kernel. One page per model. So, 3 additional pages.
8. **References**

Demo:

Your Python script.

Here are some sample datasets:

1. Flight Delays and Cancellations: https://www.kaggle.com/usdot/flight-delays
2. Heart Disease Data Set: https://archive.ics.uci.edu/ml/datasets/Heart+Disease
3. MIT Leukemia cancer dataset: http://portals.broadinstitute.org/cgi-bin/cancer/publications/pub_paper.cgi?mode=view&paper_id=43

Submission instructions:

This project has multiple due dates, tentatively:

1) March 9th: Dataset selection
2) March 30: Intro, Stats section of your individual report (words or pdf)
3) April 6: Methods and Results from Logistic regression & SVM (words or pdf)
4) April 13: Results from neural nets (words or pdf)
5) April 20: Discussions & Conclusions (words or pdf)
6) April 25: Presentation (ppt, 1 copy per group)
7) April 25: An electronic copy of your Python scripts (yes, need only 1 copy per group)
8) April 25: Final individual project report (words or pdf).