# Homework 2

Re-submit Assignment

**Due** Mar 22 by 11:59pm **Points** 100 **Submitting** a file upload (Turnitin enabled)
**File Types** pdf, txt, py, and r

**For the following assignments, please provide as much evidence of the results as possible, including the code, screenshots (only plots – not text or code) and documentation. Submit only one pdf file. All questions are equally important.**

1. Follow the simple tutorial at **https://www.machinelearningplus.com/logistic-regression-tutorial-examples-r/ (https://www.machinelearningplus.com/logistic-regression-tutorial-examples-r/)** to see Logistic Regression in action. Implement the same functionality for the same dataset in Python. Do you achieve the same accuracy as with R?

2. Choose one of the cleaned datasets at **https://www.kaggle.com/annavictoria/ml-friendly-public-datasets (https://www.kaggle.com/annavictoria/ml-friendly-public-datasets)** . Split it into training and test data. Apply any two ML algorithms that you learned in class. You can use R or Python to implement them. Which one of the algorithm fares better?

3. Refer to online tutorials on K-NN implementation such as **https://machinelearningmastery.com/tutorial-to-implement-k-nearest-neighbors-in-python-from-scratch/ (https://machinelearningmastery.com/tutorial-to-implement-k-nearest-neighbors-in-python-from-scratch/)** . Extend the implementation to use various distance metrics such as Manhattan distance and note if the classification changes with the distance metric (for an more exhaustive list of distances, see getDistMethods() in R). Use the same dataset as you used for (2) above.

4. Manually generate the decision tree (as much as possible) for the following subset from a large dataset using the ID3 algorithm. Show the information gain computation at each stage. Then generate the decision tree programmatically using R or Python. Submit code and the decision tree so generated.

| Windy? | Air Quality Good? | Hot? | Play Tennis? |
|---|---|---|---|
| No | No | No | No |

| | | | |
|---|---|---|---|
| Yes | No | Yes | Yes |
| Yes | Yes | No | Yes |
| Yes | Yes | Yes | No |

5. Listing which problem domains are best suited for each, briefly explain in your own words, the pros and cons of

- Logistic Regression
- K-NN
- SVM
- Naïve Bayes
- Decision Trees

6. Reading Assignment [Not graded as a Homework, but exams may include questions based on this assignment]: Big Data Ecosystem – study the toolset available to handle Big Data.