Assignment 2

Applied Machine Learning

1. [20 pts] At a high-level, without entering into mathematical details, compare and contrast the following classifiers:
   - Perceptron (textbook's version)
   - SVM
   - Decision Tree
   - Random Forest (you have to research a bit about this classifier)

   Some comparison criterion can be,
   - Does the method solve an optimization problem, if yes what is the cost function?
   - Speed? Strength? Robustness? Statistical?
   - Feature type that the classifier naturally uses (e.g. based on the comparison measure, such as entropy or distance)
   - Which one will be the first that you would try on your dataset?

2. [20 pts] Using real datasets (can also be hypothetically constructed by yourself) define the following feature types, and give example values from your dataset. How would you represent these features in a computer program? (e.g., 32-bit integer? Floating point? String?)
   - Numerical
   - Nominal
   - Date
   - Text
   - Image
   - Dependent variable

3. [20 pts] Using online resources, research and find other classifier performance metrics which are also as common as the accuracy metric. In your own words write down the mathematical equations and the meaning of the metrics that you found.

4. [40 pts] Implement a correlation **program from scratch** to look at the correlations between the features of `Admission_Predict_Ver1.1.csv` dataset file (not provided, you have to download it by yourself by following the instructions in the module Jupyter notebook, Graduate Admission data, 9 features, 500 data points). Display the correlation matrix where each row and column are the features, which should be an 8 by 8 matrix (should we use `'Serial no'`?). You can use pandas `DataFrame.corr()` to only verify correctness of yours. Remember, you are not allowed to used `numpy` functions such as `mean()`, `stdev()`, `cov()`, etc. except for vector/matrix arithmetic.

   Observe that the diagonal of this matrix should have all 1's and explain why? Since the last column can be used as the target (dependent) variable, what do you think about the correlations between all the variables? Which variable should be the most important for prediction of `'Chance of Admit'`?