Assignment 2

Applied Machine Learning

1. [20 pts] At a high level (i.e., without entering into mathematical details), please describe, 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:
   - Speed?
   - Strength?
   - Robustness?
   - The feature type that the classifier naturally uses (e.g. relying on distance means that numerical features are naturally used)
   - Is it statistical?
   - Does the method solve an optimization problem? If yes, what is the cost function?

   Which one will be the first that you would try on your dataset?

2. [20 pts] Define the following feature types and give example values from a dataset. You can pull examples from an existing dataset (like the Iris dataset) or you could write out a dataset yourself. (Hint: In order to give examples for each feature type, you will probably have to use more than one dataset.)
   - 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. Provide the **mathematical equations** for them *and* explain in **your own words** the meaning of the different metrics you found. Note that providing mathematical equations might involve defining some more fundamental terms, e.g. you should define "`False Positive`," if you answer with a metric that builds on that.

4. [40 pts] Implement a correlation program **from scratch** to look at the correlations between the features of `Admission_Predict.csv` dataset file. (This Graduate Admission dataset, with 9 features and 500 data points, is not provided on Canvas; you have to download it from Kaggle by following the instructions in the module Jupyter notebook.) Remember, you are not allowed to used `numpy` functions such as `mean()`, `stdev()`, `cov()`, etc.

   You may use `DataFrame.corr()` only to verify the correctness of your from-scratch matrix.

Display the correlation matrix where each row and column are the features. (Hint: this should be an 8 by 8 matrix.)

- Should we use `'Serial no'`? Why or why not?
- Observe that the diagonal of this matrix should have all 1's; why is this?
- 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 to try to predict `'Chance of Admit'`?