Voluntary Homework 1
Data Science (Summer term 2020)

Use the dataset provided along with Jupyter notebook to complete the exercise below.

Exercise 1 (Predictive modeling pipeline) *8 points*

We will use the Wisconsin Diagnostic Breast Cancer (WDBC) dataset. The dataset consists of 569 samples of biopsied tissue. The tissue for each sample is imaged and 10 characteristics of the nuclei of cells present in each image are characterized. These characteristics are Radius, Texture, Perimeter, Area, Smoothness, Compactness, Concavity, Number of concave portions of contour, Symmetry and Fractal dimension. Each of the 512 samples used in the dataset consists of a feature vector of length 30. The first 10 entries in this feature vector are the mean of the characteristics listed above for each image. The second 10 are the standard deviation and last 10 are the largest value of each of these characteristics present in each image. Each sample is also associated with a label (last column in the csv file). A label of value 1 indicates the sample was for malignant(cancerous) tissue. A label of value 0 indicates the sample was for benign tissue. The task is to create a predictive model that can classify benign and malignant tissue.

a) Split encoded dataset into training and testing parts.

b) Choose two of the classifiers mentioned in the slides from 'sklearn' python library and fit your classifiers to the given dataset. Report your accuracy, precision, recall along with confusion matrix on the test set. You can find examples here.

c) Add appropriate comments wherever necessary explaining the choice of methods you have used in your program.

Exercise 2 (Predictive modeling theory) *2 points*

1. Let's say you have achieved an accuracy score xx% depending on your choice of classifier. Do you think it is a good score? If yes, what did you compare with?