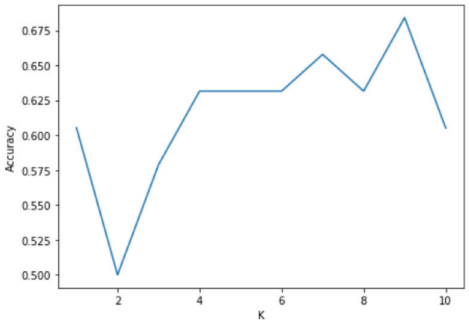
3 Nonparametric Methods (Bonus 20%)

You are given the data set of Pokemons (Pokemon.csv). This data set contains 3 classes (Water, Normal and Psychic). The column "Type 1" in the file is the labels of classes (target values), the other dimensions are the features of data. In this exercise, you will implement K-nearest-neighbor to construct a multiclass classification model. You may refer Section 2.5.

	Name	Type 1	Total	HP	Attack	Defense	Sp. Atk	Sp. Def	Speed	Generation	Legendary
0	Porygon-Z	Normal	535	85	80	70	135	75	90	4	FALSE
1	MeowsticF	Psychic	466	74	48	76	83	81	104	6	FALSE
2	Aipom	Normal	360	55	70	55	40	55	85	2	FALSE
3	Froakie	Water	314	41	56	40	62	44	71	6	FALSE
4	Slaking	Normal	670	150	160	100	95	65	100	3	FALSE
5	Slakoth	Normal	280	60	60	60	35	35	30	3	FALSE
6	Frogadier	Water	405	54	63	52	83	56	97	6	FALSE

1. K-nearest-neighbor classifier is implemented in the following procedure:
   - There are 158 data samples in this dataset.You should use first 120 samples as training data, and the remaining 38 samples as test data. (This is unbalance dataset)
   - You need to preprocess all features by subtracting the mean and normalizing by standard deviation. (formula : $\frac{x-\mu}{\sigma}$)
   - In inference stage, you compare each test sample with 120 training samples and measure the Euclidean distance between them.
   - You can use the class with the largest number of occurrences for those K closest training samples to test sample as the prediction of this test sample.
   - Try different K (from 1 to 10).
   - Plot the figure of accuracy where horizontal axis is K and vertical axis is accuracy.



2. Please implement the principal component analysis (PCA) for training samples and reduce the dimension of training and test data to 7, 6, and 5 by using the bases obtained from PCA. Repeat the above procedure.