**Introduction**

A k-nearest-neighbor (KNN) algorithm, is an approach to data classification that estimates how likely a data point is to be a member of one group or the other depending on what group the data points nearest to it are in.

KNN is an example of a "lazy learner" algorithm, meaning that it does not build a model using the training set until a query of the data set is performed.

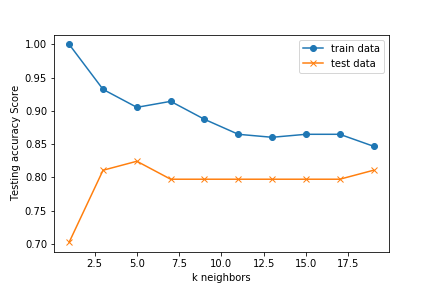
The elegance of KNN algorithm lies in its simplicity. Despite its various drawback’s such as high compute time in high dimension data it is still widely used in areas of bio-informatics and computer vision tasks.

KNN robust with regards to the search space; for instance, the classes do not have to be linearly separable. There are only two parameters to tune: k and distance metric.

**Implementation on the Heart Disease dataset**

We used KNN as a classifier algorithm. After splitting the heart disease data into train and test sets, the data was scaled and encoded. A KNN classifier model was setup, we looped through a list of K values to determine the best accuracy score.

The results showed that the K=5 results in the best score for both train and test data.



The overall accuracy of the model was 0.784.

Algorithm modification and tuning: Given the limited data samples, K-Fold Cross-Validation was used to resample and evaluate the KNN model.

In k-fold cross-validation, the original sample is randomly partitioned into k equal size sub samples. Of the k sub samples, a single sub sample is retained as the validation data for testing the model, and the remaining k-1 sub samples are used as training data. The cross-validation process is then repeated k times (the folds), with each of the k sub samples used exactly once as the validation data. The k results from the folds can then be averaged (or otherwise combined) to produce a single estimation. The advantage of this method is that all observations are used for both training and validation, and each observation is used for validation exactly once.

In addition, GridSearchCV along with cross-validation was used to evaluate the model across various parameters.

The graph below shows mean accuracy across the different k values after the k-fold cross-validation and GridSearchCV was used.

