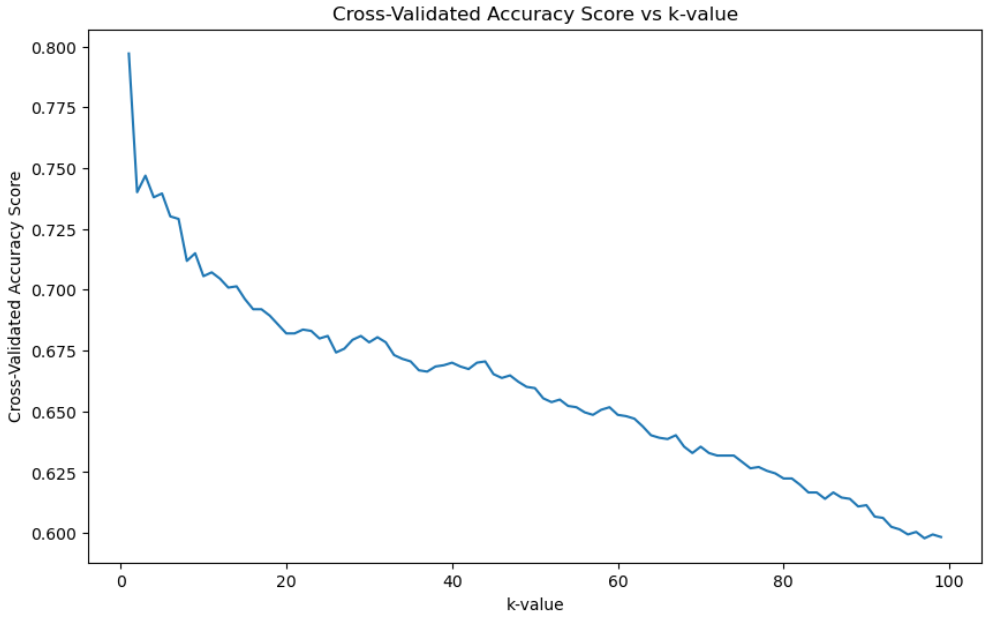
02-725 HW6

Anushka Sinha

1 SUBCELLULAR LOCALIZATION

1. Refer to the code to see which features are discrete or Gaussian. The accuracy of this method is about 0.614.

2. Here is the elbow plot obtained by trying a range of values for k. The features were first scaled using StandardScaler(). The accuracies in the plot was calculated using training data and 5 fold cross-validation. The value of k = 1 obtained the highest accuracy of 0.797. However, the high accuracy at k = 1 might be due to over-fitting. According to the elbow plot, k = 6 is the optimal k value for reasonable accuracy score and lower chances of over-fitting.



3. Bottlenecks of Naive Bayes Classification:
•It assumes that all features are independent of each other which is rarely the case in real life.
•Since it assigns a probability of 0 to a categorical variable whose category in the test data wasn't available in the training dataset, it leads to the zero-frequency problem.

Bottlenecks of kNN classifier:
•Since the classifications produced by the kNN classifier are based entirely on the local data structure, the algorithm is very sensitive to the noise in the data.
•kNN classifier is not suitable for large-dimensional datasets.
•The performance of the classifier is heavily dependent on the value of k chosen.
•Since kNN does not build any traditional machine learning model and makes predictions using the entire dataset, it is not able to recognize the underlying patterns in the data.