Claude Shyaka
ID#: 801326243
ECGR-5105

Homework 3: Naïve Bayes and PCA Feature Extraction

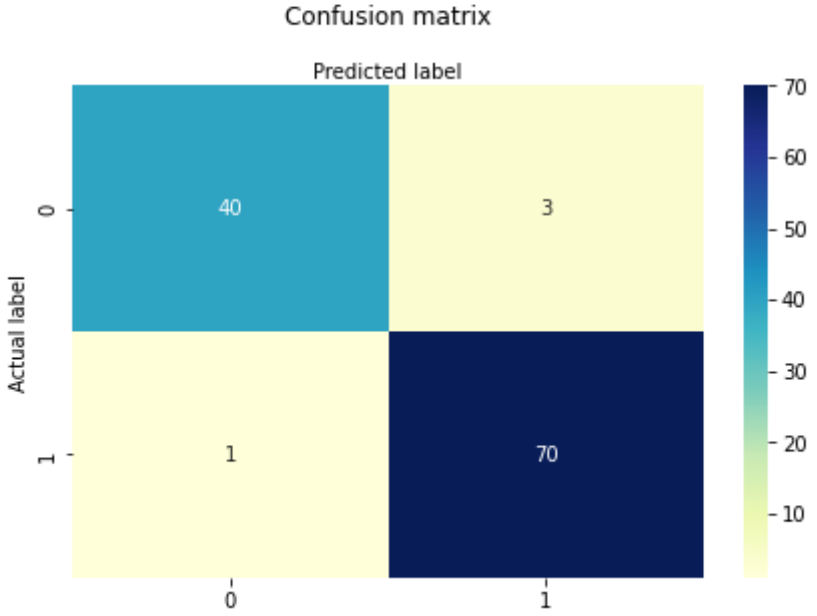
GitHub Repo: https://github.com/claudeshyaka/ml

In this exercise, we used the cancer dataset to build a Naïve Bayesian model to classify the type of cancer (Malignant vs. benign). In addition, PCA feature extraction was used to reduce the complexity of our dataset. The dataset was split into 80% for training and 20% for testing. In addition, both train and test sets were scaled using the Standard Scaler from Sklearn. The following results were obtained:

1. A Naïve Bayes classifier was trained with the cancer dataset and the following results were obtained:

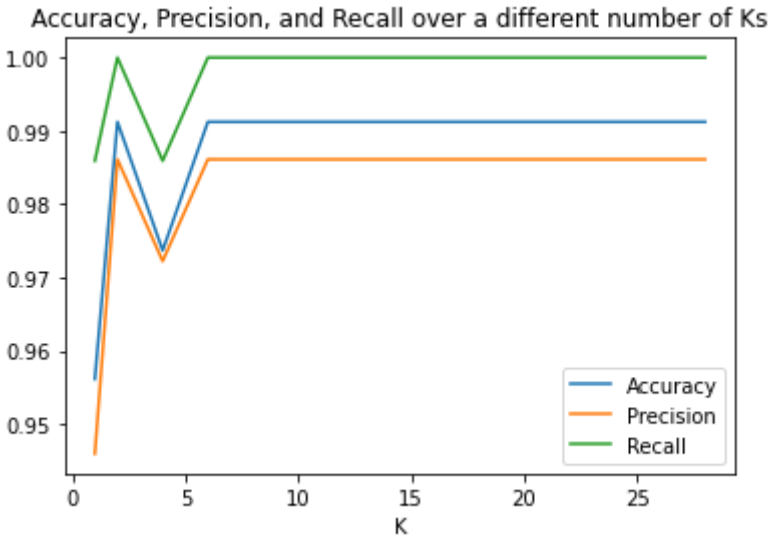
Classification metrics:

	precision	recall	f1-score	support
0	0.98	0.93	0.95	43
1	0.96	0.99	0.97	71
accuracy			0.96	114
macro avg	0.97	0.96	0.96	114
weighted avg	0.97	0.96	0.96	114



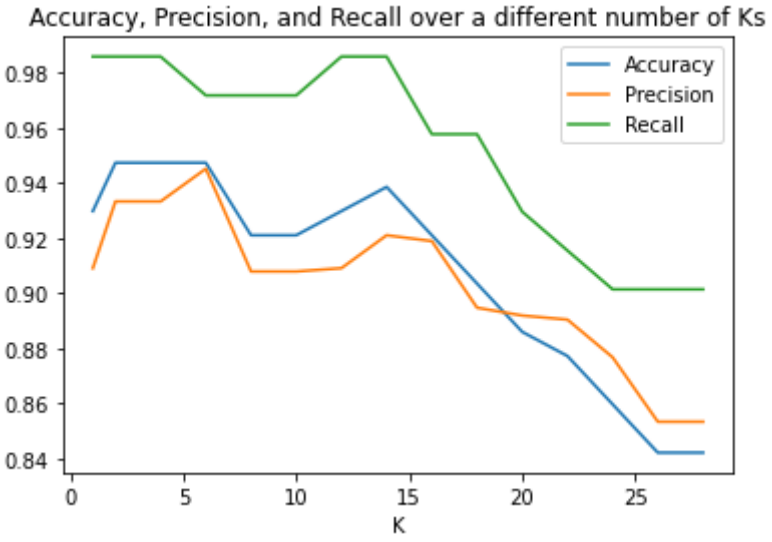
Comparing results from here with those obtained in the previous homework, we can see that the Naïve Bayes classifier has a lower accuracy compared to the Logistic regression classifier from Homework 2.

2. This section introduced the PCA feature extraction and trained a logistic regression classifier with features extracted by the PCA feature extraction. The following results were obtained for accuracy, precision, and recall for different k values.



As shown in the figure above, a value of 5 for k returns the optimal accuracy, precision, and recall. In addition, after reaching the optimal value of 5 for k, increasing the number of features has no effect on the overall accuracy, precision, and recall of the classifier.

3. This section combines PCA features extraction with a Naïve Bayes classifier. The following results were obtained for accuracy, precision, and recall for different k values.



As shown in the figure above, as the number of features increases, the accuracy, precision, and recall keep decreasing. These results affirm our understanding of the Naïve Bayes classifier in that it is a classifier that works best with data that contains fewer input features whereas, in problem 2, the logistic regression classifier performed well even after adding more features to the dataset.