**Homework 3**

GitHub Repository Link: <https://github.com/NaraPvP/IntroToML>

For all the problems, an 80/20 training split was done on the dataset. Along with this, the features were scaled using standardization.

**Problem 1:**

For the Naïve Bayesian model, the following was the classification report (which provides the accuracy, precision, and recall):

Table

Description automatically generated with medium confidence

Along with this, the confusion matrix is shown here for the Naïve Bayesian model: 

With the logistic regression classifier that was used in the last homework, this was the classification report and confusion matrix:

Table

Description automatically generated

When comparing these two classifiers, all of the classification parameters (accuracy, precision, and recall) are better for the logistic regression model. This implies that the logistic regression model is better for the cancer dataset compared to the Naïve Bayesian model.

**Problem 2:**

After fitting the PCA feature extraction with the training set of our features, I applied the PCA feature extraction on the test set of features. Afterwards, a logistic regression model was trained using the PCA-adjust features as new training and test set. By running the following code, it was determined the optimum number of K in a general case is 9 principal components:

Graphical user interface, application

Description automatically generated

For the case shown in the code PDF, 4 principal components gave the highest classification accuracy. The plot of the classification report will be provided here:

Chart, line chart

Description automatically generated

The accuracy is the highest statistic from the classification report, which is plausible due to more predicted values being able to seem accurate as discussed in lecture. We want our precision to be high due to the nature of the PCA feature extraction, which desires to keep as much information from the original set as possible.

**Problem 3:**

Similarly to Problem 2, the PCA feature extraction was used for training a Naïve Bayes model. The plot of the classification report as k (number of principal components) increase is shown here:

Chart, line chart

Description automatically generated

Like the end of Problem 1, the Bayes classifier does not perform as well as the Logistic Regression classifier. This was to be expected as both datasets had changed equally since Problem 1. Since Problem 2 (Logistic Regression model) had higher statistics overall, this further reinforces that a Logistic Regression classifier is more accurate than a Naïve Bayes classifier.