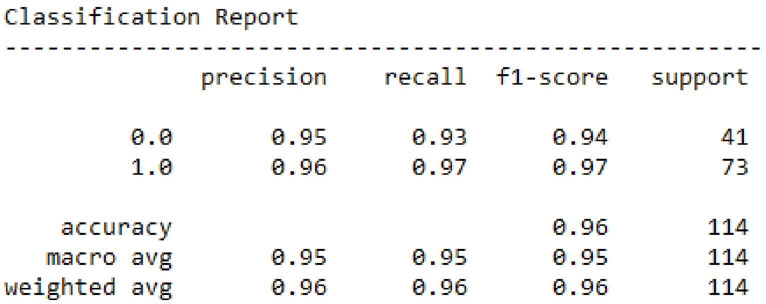
GitHub Code Repository: https://github.com/pharmon0/Harmon_ECGR5105

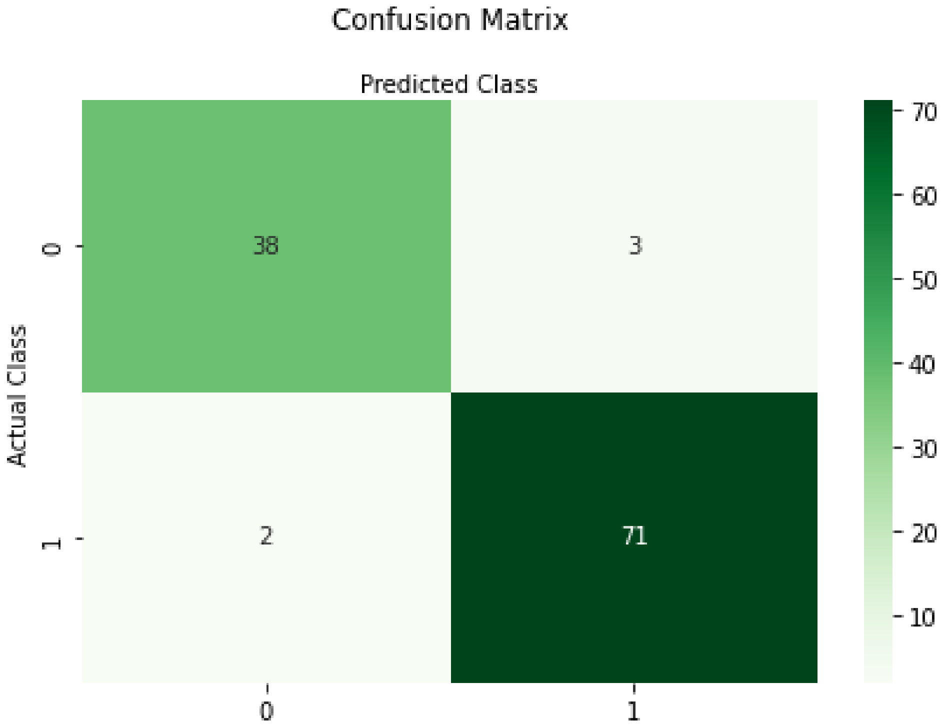
Problem 1)

Use the cancer dataset to build a Naïve Bayesian model to classify the type of cancer (Malignant vs. benign). Plot your classification accuracy, precision, and recall. Explain and elaborate on your results. Can you compare your results against the logistic regression classifier you did in previous homework.

Using the Gaussian Naïve Bayes classifier, the following results were achieved with the cancer dataset:

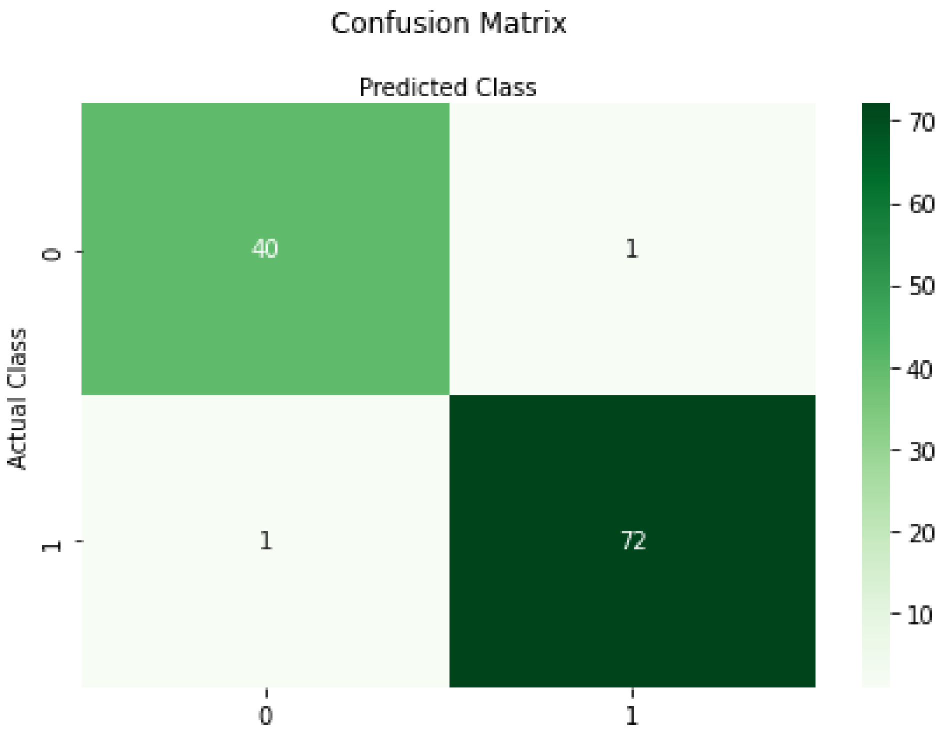


```
Classification Report
-----------------------------------------------------
              precision    recall  f1-score   support

         0.0       0.95      0.93      0.94        41
         1.0       0.96      0.97      0.97        73

    accuracy                           0.96       114
   macro avg       0.95      0.95      0.95       114
weighted avg       0.96      0.96      0.96       114
```



The results of the logistic regression problem as shown in Homework 2 were:

```
Model Accuracy:  98.246%
Model Precision: 98.630%
Model Recall:    98.630%
```

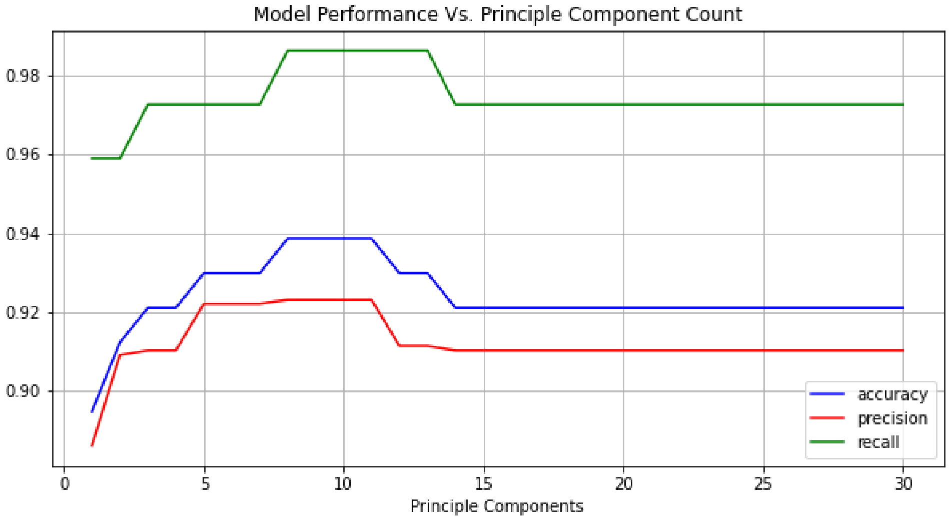


From this, it can be seen that the logistic regression model from Homework 2 held higher accuracy, recall, and precision, correctly identifying more true positives and more true negatives. The logistic regression yielded a roughly 2% gain in accuracy over the Gaussian Naïve Bayes model. This confirms what was expected: It is easier to get a good accuracy on a discriminative model like logistic regression over a generative model like Naïve Bayes.

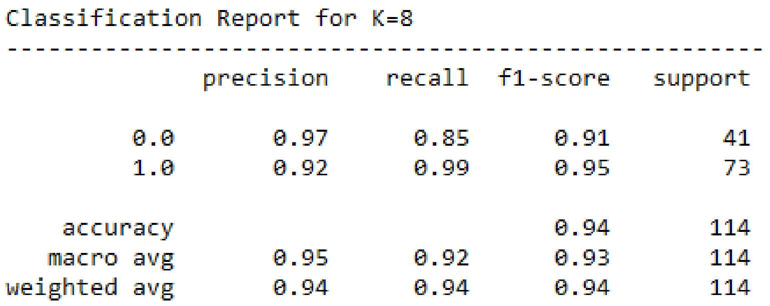
Problem 2)

Use the cancer dataset to build a logistic regression model to classify the type of cancer (Malignant vs. benign). Use the PCA feature extraction for your training. Perform N number of independent training (N=1, …, K). Identify the optimum number of K, principal components that achieve the highest classification accuracy. Plot your classification accuracy, precision, and recall over a different number of Ks. Explain and elaborate on your results.

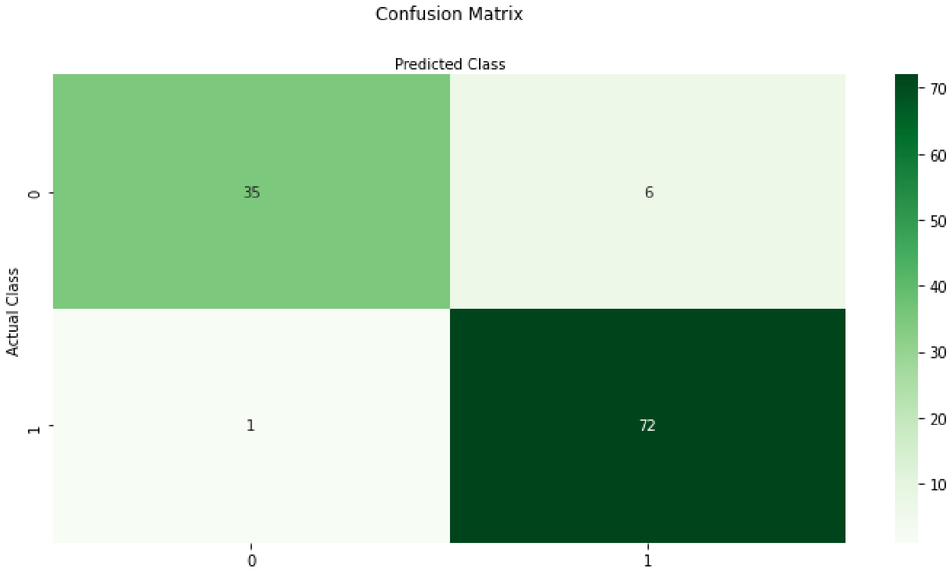
The logistic regression training was run for a wide range of PCA K-values to find the peak accuracy using the smallest number of principle components. A chart of model performance vs. K can be seen below.



As can be seen in the chart above, the peak accuracy occurs at K=8. The results of this model can be seen below.



```
Classification Report for K=8
---------------------------------------------------
              precision    recall  f1-score   support

         0.0       0.97      0.85      0.91        41
         1.0       0.92      0.99      0.95        73

    accuracy                           0.94       114
   macro avg       0.95      0.92      0.93       114
weighted avg       0.94      0.94      0.94       114
```

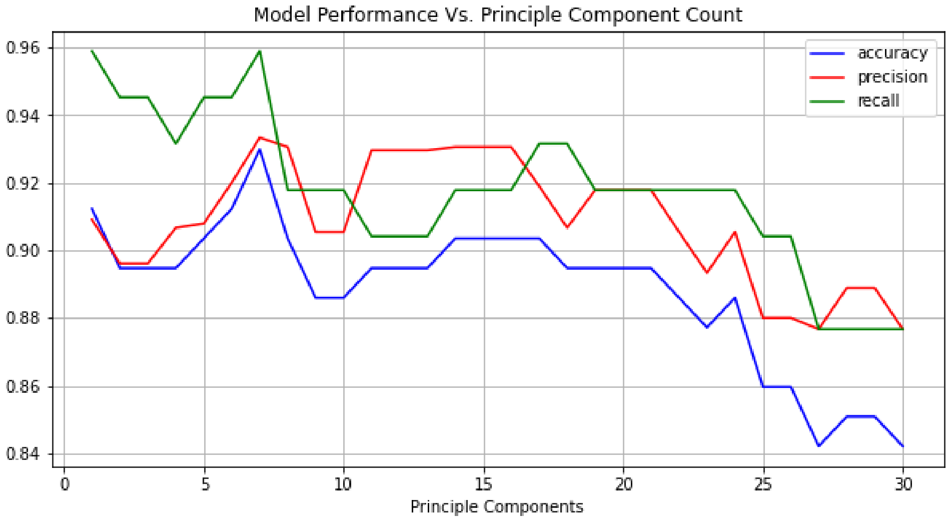


Unfortunately, in this case, it seems that the PCA feature reduction has had a negative impact on the whole logistic regression model performance. No performance metric of the PCA feature reduced model seems to be any significant improvement on the logistic regression model from Homework 2.

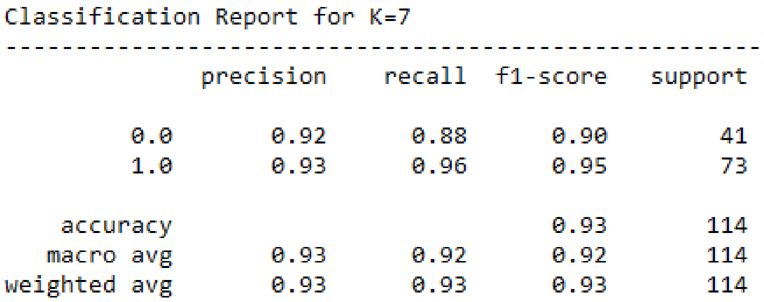
Problem 3)

Can you repeat problem 2? This time, replace logistic regression with the Bayes classifier. Report your results (classification accuracy, precision, and recall). Compare your results against problem 2.

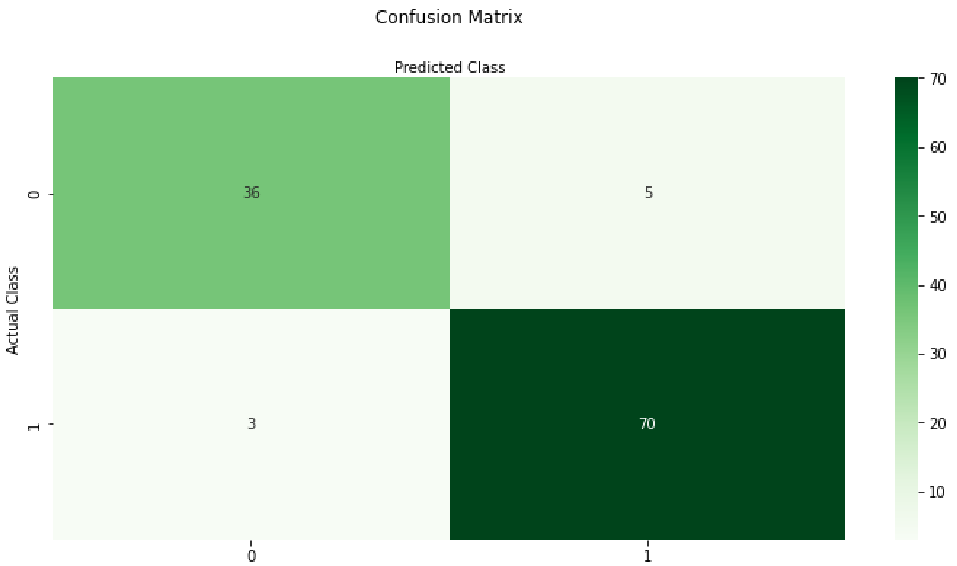
The Naive Bayes training was run for a wide range of PCA K-values to find the peak accuracy using the smallest number of principle components. A chart of model performance vs. K can be seen below.



As can be seen in the chart above, the peak accuracy occurs at K=7. The results of this model can be seen below.



```
Classification Report for K=7
-----------------------------------------------------
              precision    recall  f1-score   support

         0.0       0.92      0.88      0.90        41
         1.0       0.93      0.96      0.95        73

    accuracy                           0.93       114
   macro avg       0.93      0.92      0.92       114
weighted avg       0.93      0.93      0.93       114
```



Unfortunately, in this case, it seems that the PCA feature reduction has had a negative impact on all model performance parameters of the Naïve Bayes model from Problem 1. There is a notable reduction in recall, precision, and accuracy.

On the upside, it appears that the precision score of this model is slightly better than the logistic regression of problem 2. All other performance characteristics seem to be worse overall.