ECGR 4105: Intro to Machine Learning HW 3

Axel Leon Vasquez

Student ID: 801182414

October 15th, 2024

Link to GitHub

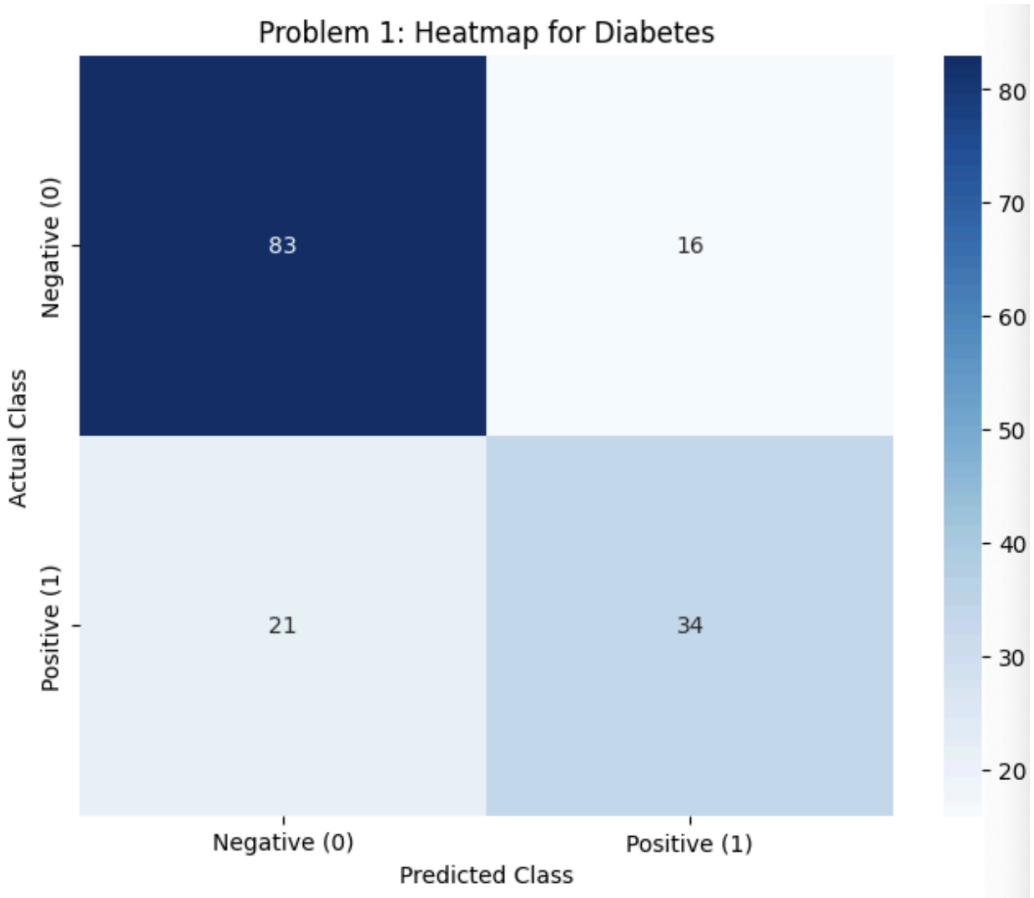
Problem 1

1. **Introduction:**
   This task aimed to build a logistic regression binary classifier to predict the likelihood of diabetes based on a given set of input variables including:
   - Pregnancies
   - Glucose
   - Blood Pressure
   - Skin Thickness
   - Insulin
   - BMI
   - Diabetes Pedigree Function
   - Age

2. **Methodology:**
   The dataset input was split into 80% for training and 20% for testing. The input features were standardized to ensure all variables were on a similar scale. This scaling was applied to both the training and testing datasets. A logistic regression model was built using LogisticRegression() with the liblinear solver and fit to the training data. The model predicted the outcomes on the test data, and the predicted probabilities of positive outcomes were also calculated.

3. **Results:**
   The following evaluation metrics were derived from the model's performance on the test set:
   - Accuracy: 0.76 (76%)
   - Precision: 0.68
   - Recall: 0.62
   - F1-Score: 0.65

   A confusion matrix was plotted to visualize the classification results, where the heatmap displayed the number of true positives, true negatives, and false positives.



4. **Summary of Findings**
   The logistic regression model achieved an accuracy of 76%, indicating a solid performance in predicting diabetes. The precision of 0.68 suggests the model is reasonably good at identifying true positives (diabetics) while minimizing false positives. However, the recall score of 0.62 shows that the model misses some diabetic cases, as it does not correctly identify all true positives. The confusion matrix shows that 83 out of 99 non-diabetic cases were correctly classified (true negatives), while 34 out of 55 diabetic cases were correctly classified (true positives). However, there were 21 false negatives, meaning 21 diabetic cases were misclassified as non-diabetic. This reveals some limitations in the model's ability to capture diabetic cases fully.

Problem 2 - Part 1

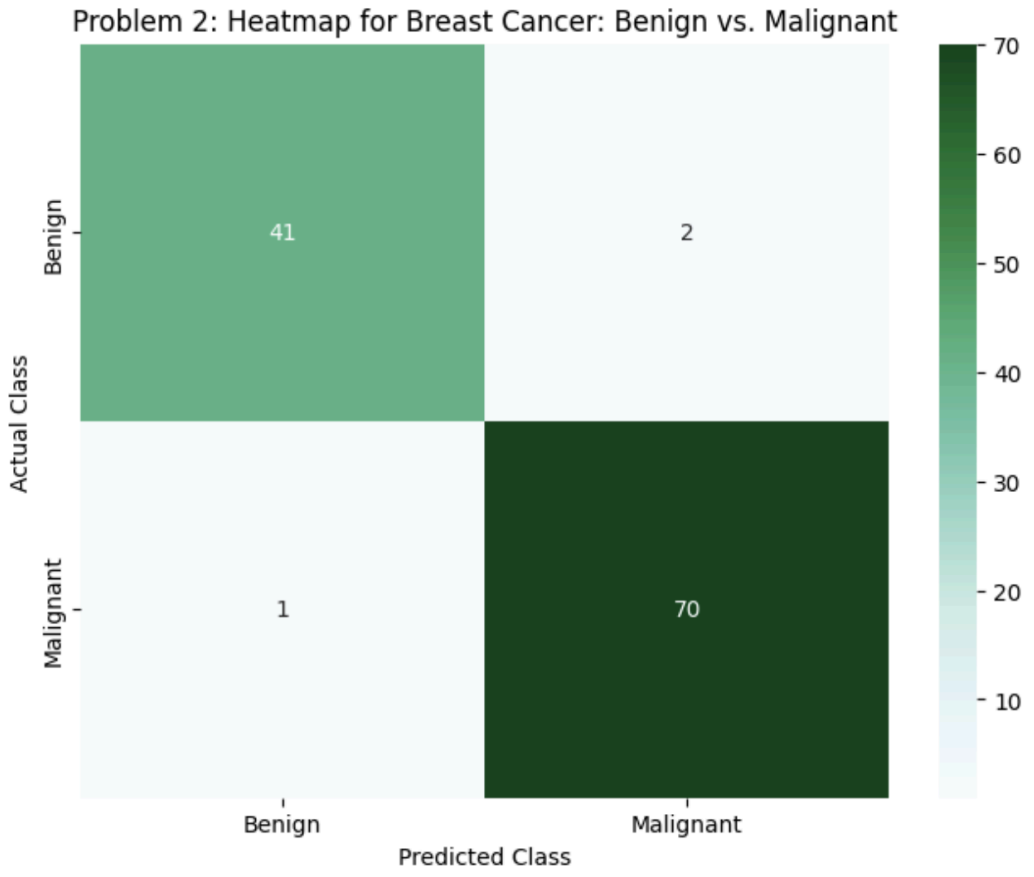
1. **Introduction:**
   This task focused on analyzing a breast cancer dataset to classify tumors as either benign or malignant. The dataset used for this classification was the Breast Cancer Wisconsin (Diagnostic) Dataset, which is commonly used in machine learning for binary classification tasks. The dataset includes 30 features derived from the characteristics of cell nuclei present in digitized images of a breast mass.
   - **Input Variables:** 30 numeric features representing properties such as radius, texture, perimeter, area, and smoothness of the tumor cells.
   - **Output Variable:** A binary classification where:

- 0 represents benign
- 1 represents malignant

2. **Methodology:**
   The dataset input was split into 80% training and 20% testing. The input features were standardized and scaled to ensure all variables were on a similar scale. A logistic regression model was developed with a maximum iteration of 10,000 to ensure proper convergence. The model was used to predict the classes for the test set. Additionally, the probability of the positive class was calculated.

3. **Results:**
   The logistic regression model produced the following results:
   - Accuracy: 0.9737 (97.37%)
   - Precision: 0.9722
   - Recall: 0.9859
   - F1-Score: 0.9790

   A confusion matrix was plotted to visualize the classification results, where the heatmap displayed the number of true positives, true negatives, and false positives.



The model achieved a high overall accuracy of 97.37%, with a precision of 0.9722, indicating that when the model predicts a tumor as malignant, it is correct 97.22% of the time. The recall of 0.9859 shows that the model successfully identified 98.59% of malignant cases, while the F1-score of 0.9790 indicates a good balance between precision and recall.

4. **Summary of Findings**
   The logistic regression model effectively classified breast tumors as either benign or malignant, achieving an accuracy of 97.37%. The model's high precision (97.22%) and recall (98.59%) demonstrate its ability to minimize both false positives and false negatives. The confusion matrix shows that only 2 benign tumors were misclassified as malignant, and only 1 malignant tumor was misclassified as benign. Overall, the model's strong performance suggests that logistic regression is a suitable choice for this binary classification problem, offering reliable predictions with minimal error.

Problem 2 - Part 2

1. **Introduction:**
   This task builds on the logistic regression model from Problem 2 Part 1 by exploring different values of the regularization parameter C to determine its impact on the model's performance. The C parameter controls the regularization strength in logistic regression, with smaller values indicating stronger regularization. By varying the C values, the model aims to identify the optimal regularization that balances model complexity and performance, preventing both overfitting and underfitting.

2. **Methodology:**
   The C parameter was tested over a range of values [0.001, 0.01, 0.1, 1, 10]. A smaller value of C applies stronger regularization, encouraging the model to simplify, while larger values allow more complexity in fitting the data.

   For each value of C, a logistic regression model was trained on the standardized training set and then used to predict outcomes on the test set. Key performance metrics such as: accuracy, precision, recall, and F1-score were computed to evaluate the model for each value of C.

3. **Results:**
   The following table summarizes the performance of the logistic regression model for different values of C:

C Value	Accuracy	Precision	Recall	F1-Score
0.001	0.88596491	0.84523810	1.00000000	0.91612903
0.01	0.96491228	0.94666667	1.00000000	0.97260274
0.1	0.98245614	0.97260274	1.00000000	0.98611111

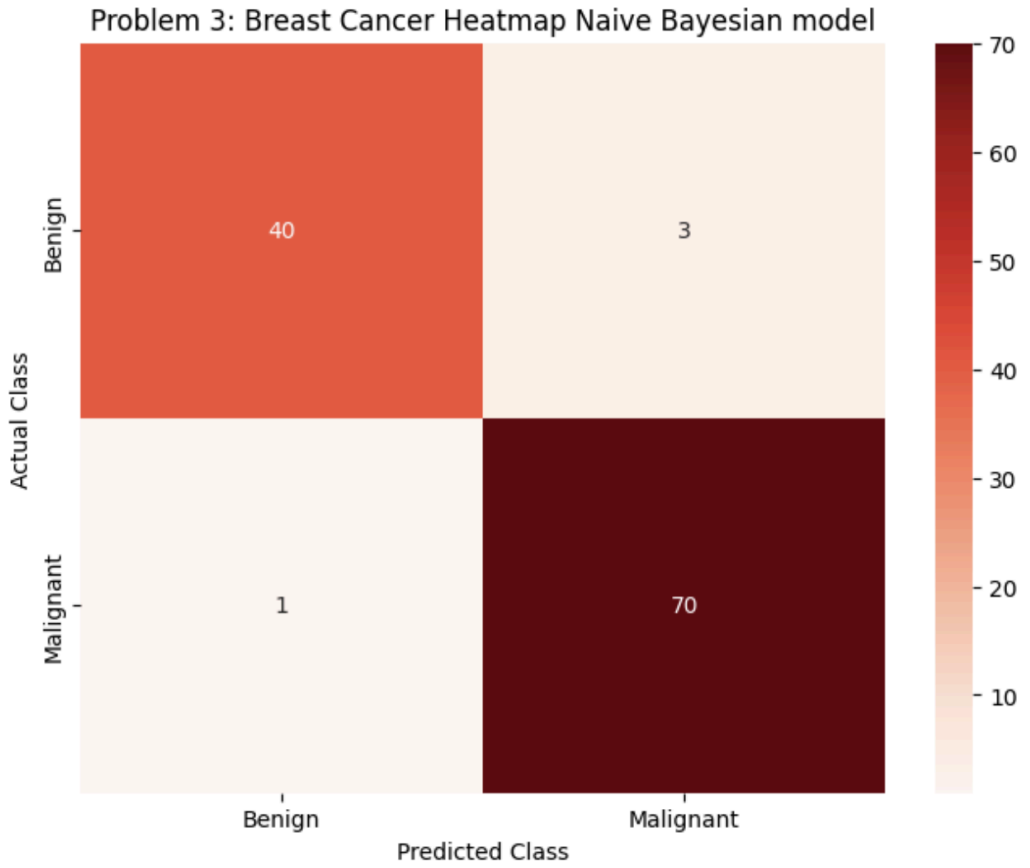
1	0.97368421	0.97222222	0.98591549	0.97902098
10	0.97368421	0.98571429	0.97183099	0.97872340

4. **Summary of Findings**
   Based on the performance metrics for various values of C, C = 0.1 emerged as the optimal choice for this model. It provides the best balance across all metrics, with the highest F1-score of 98.61%, indicating that the model generalizes well without overfitting or underfitting. This value of C allows the model to effectively distinguish between benign and malignant tumors, minimizing false positives and negatives.

Problem 3

1. **Introduction:**
   The Naive Bayes classifier was used to train the breast cancer dataset in this task. Naive Bayes is a probabilistic classifier based on Bayes' theorem, assuming that the features are conditionally independent given the class label. This approach is particularly effective when dealing with high-dimensional data. The goal is to evaluate the classifier's ability to distinguish between benign and malignant tumors using the standardized dataset.

2. **Methodology:**
   Similarly to how problem 2 part 1 is set up, problem 3 has the same setup but is given a Gaussian Naive Bayes classifier and the model was trained on the scaled version of the dataset. The classifier was trained using the standardized training data, and predictions were made on the test data.
3. **Results:**
   - Accuracy: 0.9649 (96.49%)
   - Precision: 0.9589
   - Recall: 0.9859
   - F1 Score: 0.9722



4. **Naive Bayes vs. Logistic Regression Comparison**

Metric	**Naive Bayes**	**Logistic Regression (Untuned)**	**Logistic Regression (Tuned, C = 0.1)**
Accuracy	96.49%	97.37%	98.25%
Precision	95.89%	97.22%	97.26%
Recall	98.59%	98.59%	100%
F1 Score	97.22%	97.90%	98.61%
False Positive	3	2	2
False Negatives	1	1	0

Conclusion:
In summary, while both models performed well, the tuned logistic regression model (with C=0.1) showed the best results overall, achieving the highest accuracy (98.25%) and perfect recall (100%). Naive Bayes also performed competitively but slightly lagged in precision and F1 score.

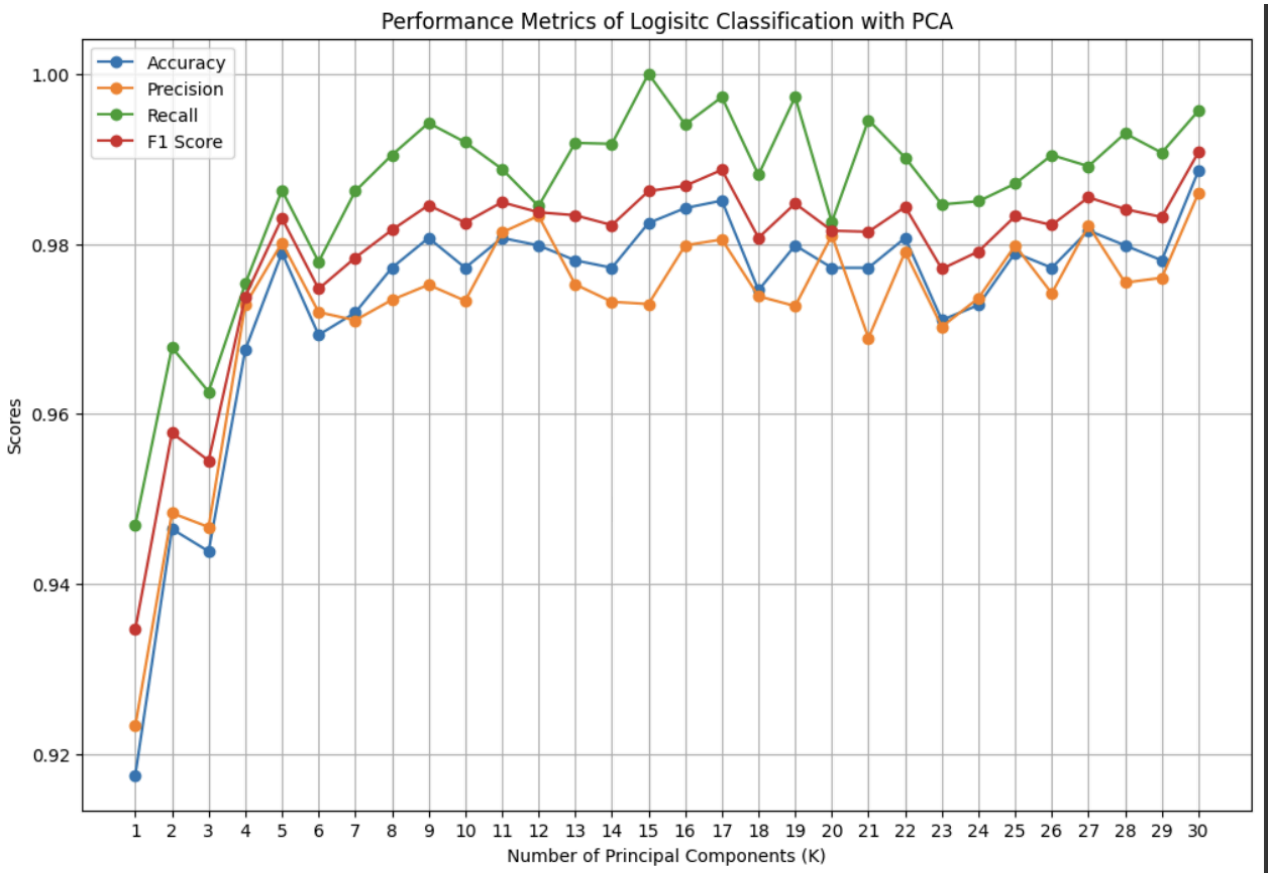
Logistic regression, especially when tuned, demonstrated better generalization and fewer classification errors, making it the more effective model for this dataset.

Problem 4

1. **Introduction:**
   This problem utilizes the same cancer dataset to build a logistic regression model aimed at classifying tumors as either malignant or benign. The primary focus is on implementing Principal Component Analysis (PCA) for feature extraction and dimensionality reduction. By examining different values of K (the number of principal components), the goal is to determine the optimum number that maximizes the model's classification performance.

2. **Methodology:**
   PCA was performed to reduce the dimensionality of the dataset. Values of K ranging from 1 to 20 were tested to find the optimal number of principal components. 10 independent training sessions were conducted for each K value. In each session, the database was split into 80% training and 20% testing. A logistic regression model was trained on the training set. The same performance metrics were calculated for each session.

3. **Results:**
   The analysis revealed that the optimal number of principal components K was 30, achieving the highest classification accuracy.



4. **Logistic Regression w/ PCA vs. Problem 2 Classifiers Comparison**
   The use of PCA in Problem 4 resulted in notable improvements across all performance metrics when compared to the standard logistic regression model in Problem 2. This indicates that dimensionality reduction via PCA not only boosts the model's performance but also helps mitigate overfitting, leading to a more reliable classification of cancer types. The optimal number of principal components, K=30, highlights the model's ability to preserve essential features of the dataset while attaining high accuracy and dependability.

Metric	**Problem 2 - Part 1** (Logistic Regression)	**Problem 2 - Part 2** (Logistic Regression w/ C = 0.10)	**Problem 4** (Logistic Regression w/ PCA) Optimum K = 30
Accuracy	0.9737	0.9825	0.9886
Precision	0.9722	0.9726	0.9860
Recall	0.9859	1.0000	0.9956
F1 Score	0.9790	0.9861	0.9908

5. **Logistic Regression w/ PCA vs. Problem 3 Classifiers Comparison**
   The comparison between Problem 3 and Problem 4 reveals significant enhancements in performance metrics through the application of PCA. The logistic regression model in Problem 4 achieved a higher accuracy, surpassing the 0.9649 accuracy of the Naive Bayes classifier in Problem 3. Additionally, the precision, recall, and F1 scores improved dramatically. These results demonstrate that the use of PCA not only enhances classification performance but also provides a more reliable model for distinguishing between cancer types.

Metric	**Problem 3** (Naive Bayes)	**Problem 4** (Logistic Regression w/ PCA) Optimum K = 30
Accuracy	0.9649	0.9886
Precision	0.9589	0.9860
Recall	0.9859	0.9956
F1 Score	0.9722	0.9908

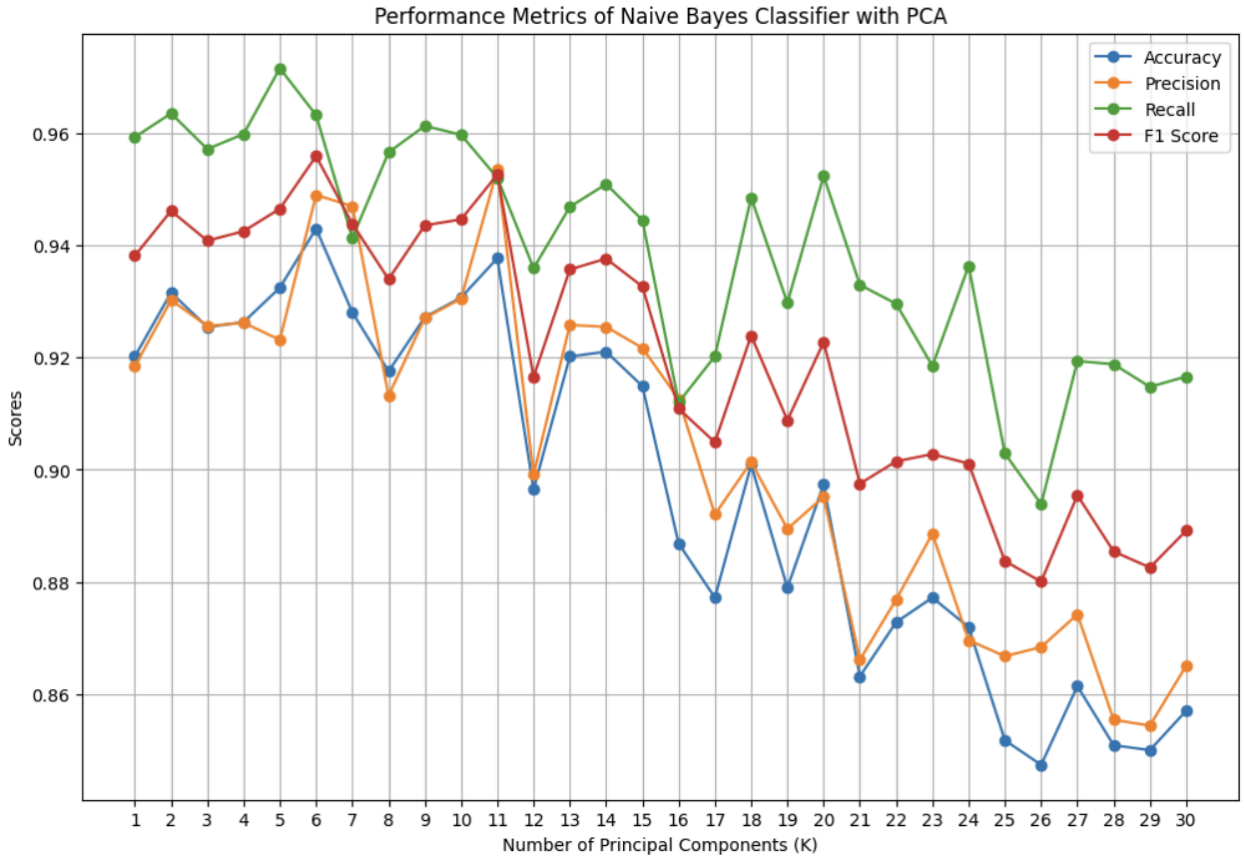
Problem 5

1. **Introduction:**
   This problem focuses on applying the naive Bayes classifier to the breast cancer dataset for classifying tumors as malignant or benign. The approach incorporates Principal Component

Analysis (PCA) for dimensionality reduction, aiming to determine the optimal number of principal components that maximize classification performance.

2. **Methodology:**
   The model was tested with K values ranging from 1 to 30 principle components to identify the optimal number of components. Similar to problem 4, 10 independent training sessions were made for each K value. A naive Bayes classifier was trained on the training set instead. By examining different values of K (the number of principal components), the goal is to determine the optimum number that maximizes the model's classification performance.

3. **Results:**
   The analysis revealed that the optimal number of principal components K was 6, which resulted in the highest classification performance



4. **Naive Bayes Classifier with PCA vs. Problem 2 Classifier**
   The comparison of performance metrics across different models highlights that Problem 2 yielded better results than Problem 5. The logistic regression models in Problem 2 demonstrated superior performance in terms of accuracy compared to the Naive Bayes classifier with PCA used in Problem 5. However, the Naive Bayes model, despite its lower accuracy, still showed promising precision, recall, and F1 scores, indicating a balanced ability to identify positive instances. This suggests that while logistic regression may provide higher accuracy, the application of PCA with Naive Bayes can enhance other important performance metrics, making it a valuable approach in

specific classification scenarios. The optimal number of principal components in Problem 5, K=6, indicates that the model successfully captures key features of the dataset while maintaining reliability in its classifications.

Metric	**Problem 2 - Part 1** (Logistic Regression)	**Problem 2 - Part 2** (Logistic Regression w/ C = 0.10)	**Problem 5** (Naive Bayes w/ PCA) Optimum K = 6
Accuracy	0.9737	0.9825	0.9430
Precision	0.9722	0.9726	0.9490
Recall	0.9859	1.0000	0.9632
F1 Score	0.9790	0.9861	0.9559

5. **Naive Bayes Classifier with PCA vs. Problem 3 Classifier**
   The comparison of performance metrics across different models highlights that Problem 2 yielded better results than Problem 5. The logistic regression models in Problem 2 demonstrated superior performance in terms of accuracy compared to the Naive Bayes classifier with PCA used in Problem 5. However, the Naive Bayes model, despite its lower accuracy, still showed promising precision, recall, and F1 scores, indicating a balanced ability to identify positive instances. This suggests that while logistic regression may provide higher accuracy, the application of PCA with Naive Bayes can enhance other important performance metrics, making it a valuable approach in specific classification scenarios. The optimal number of principal components in Problem 5, K=6, indicates that the model successfully captures key features of the dataset while maintaining reliability in its classifications.

Metric	**Problem 3** (Naive Bayes)	**Problem 5** (Naive Bayes w/ PCA) Optimum K = 6
Accuracy	0.9649	0.9430
Precision	0.9589	0.9490
Recall	0.9859	0.9632
F1 Score	0.9722	0.9559

6. **Naive Bayes Classifier with PCA vs. Problem 4 Classifier**
   The performance comparison between Problem 4 (Logistic Regression with PCA) and Problem 5 (Naive Bayes with PCA) shows that Logistic Regression outperforms Naive Bayes across all metrics. Logistic Regression achieves higher accuracy, precision, recall, and F1 score compared to Naive Bayes. While Naive Bayes maintains a decent level of sensitivity, it falls short overall,

indicating that Logistic Regression with PCA is the more effective model for classifying cancer types based on the given data.

Metric	**Problem 4** (Logistic Regression w/ PCA) **Optimum K = 30**	**Problem 5** (Naive Bayes w/ PCA) **Optimum K = 6**
Accuracy	0.9886	0.9430
Precision	0.9860	0.9490
Recall	0.9956	0.9632
F1 Score	0.9908	0.9559