Name: Abdirahim Ahmed
Student ID: 801338570

Githubrepository:https://github.com/Abdiirahim/ECGR-4105-Intro-to-ML/tree/main/Homewor3

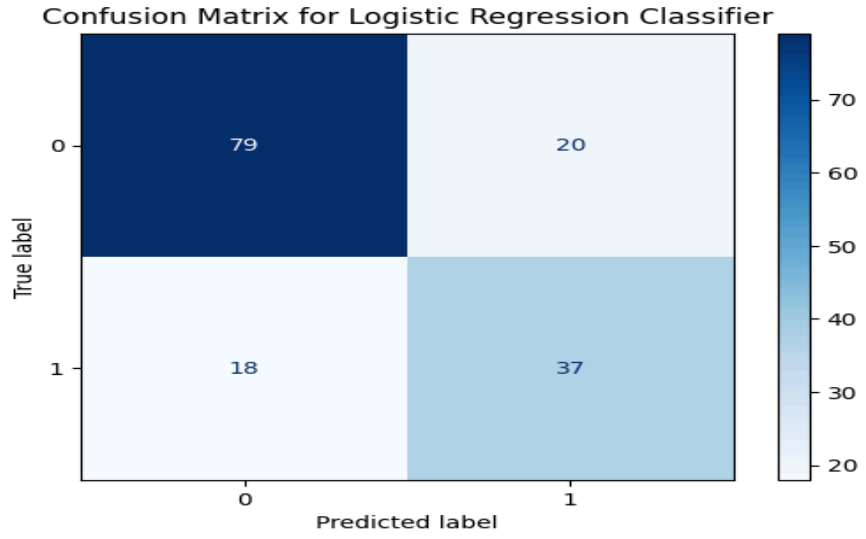
Homework 3

1. Using the diabetes dataset, build a logistic regression binary classifier for positive diabetes. Please use 80% and 20% split between training and evaluation (test). Make sure to perform proper scaling and standardization before your training. Draw your training results, including loss and classification accuracy over iterations. Also, report your results, including accuracy, precision, and recall, FI score. At the end, plot the confusion matrix representing your binary classifier.

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	\
0	6	148	72	35	0	33.6	
1	1	85	66	29	0	26.6	
2	8	183	64	0	0	23.3	
3	1	89	66	23	94	28.1	
4	0	137	40	35	168	43.1	

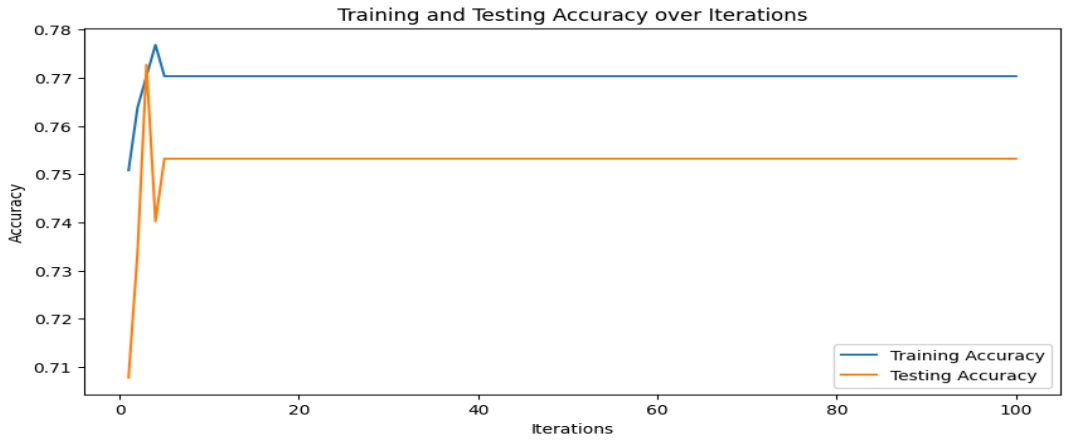
	DiabetesPedigreeFunction	Age	Outcome
0	0.627	50	1
1	0.351	31	0
2	0.672	32	1
3	0.167	21	0
4	2.288	33	1

Model Evaluation Metrics:
Accuracy: 0.75
Precision: 0.65
Recall: 0.67
F1 Score: 0.66



Name: Abdirahim Ahmed
Student ID: 801338570

Confusion Matrix Plot: A confusion matrix will be displayed, showing the true positives, true negatives, false positives, and false negatives.



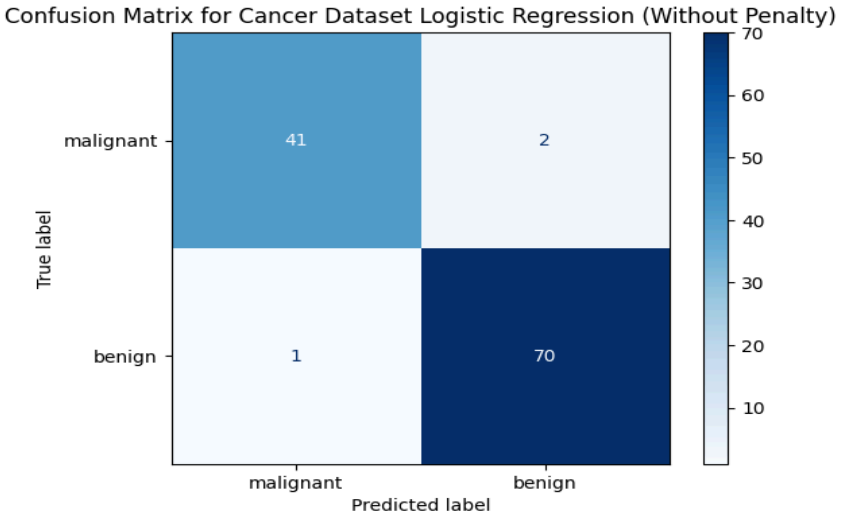
Training and Testing Accuracy Plot: A graph displaying the training and testing accuracy over the 100 iterations.

The logistic regression binary classifier achieved moderate performance. The confusion matrix plot indicates a balance between correctly identified positive and negative cases, though there is room for improvement in precision and recall. The accuracy and F1 score suggest that the model can reasonably distinguish between positive and negative diabetes cases

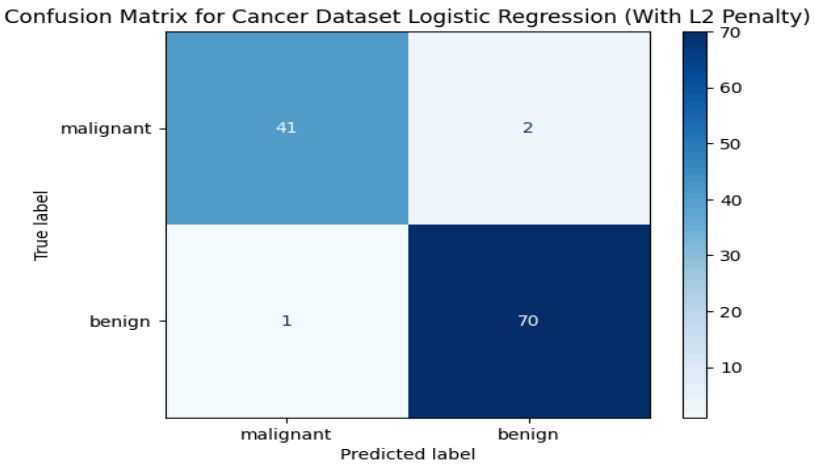
2. Use the cancer dataset to build a logistic regression model to classify the type of cancer (Malignant vs. benign). First, create a logistic regression that takes all 30 input features for classification. Please use 80% and 20% split between training and evaluation (test). Make sure to perform proper scaling and standardization before your training. Also, report your results, including accuracy, precision, recall and F1 score. At the end, plot the confusion matrix representing your binary classifier. How about adding a weight penalty here, considering the number of parameters. Add the weight penalty and repeat the training and report the results.

Cancer Dataset Logistic Regression Metrics (Without Penalty):
Accuracy: 0.97
Precision: 0.97
Recall: 0.99F1 Score: 0.98

Name: Abdirahim Ahmed
Student ID: 801338570



Cancer Dataset Logistic Regression Metrics (With L2 Penalty):
Accuracy: 0.97
Precision: 0.97
Recall: 0.99
F1 Score: 0.98



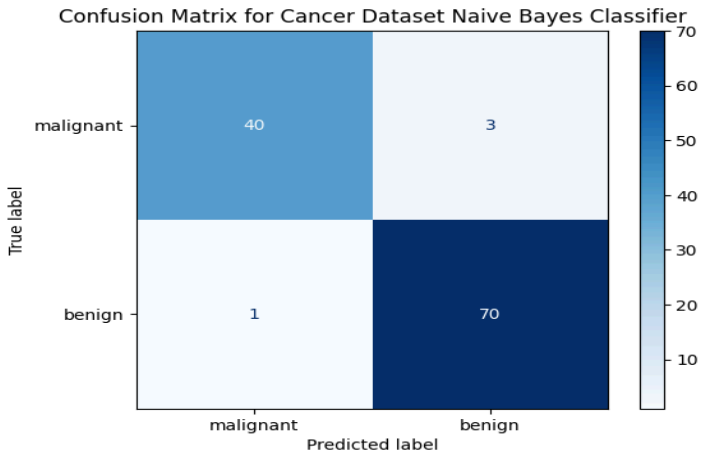
Both models (with and without penalty) performed exceptionally well, reflecting the high separability of benign and malignant cases in the dataset. The penalty did not significantly impact the performance, likely because the original logistic regression model was already well-tuned for this dataset.

3.Use the cancer dataset to build a naive Bayesian model to classify the type of cancer (Malignant vs. benign). Use 80% and 20% split between training and evaluation (test). Plot your

Name: Abdirahim Ahmed
Student ID: 801338570

classification accuracy, precision, recall, and F1 score. Explain and elaborate on your results. Can you compare your results against the logistic regression classifier you did in Problem 2.

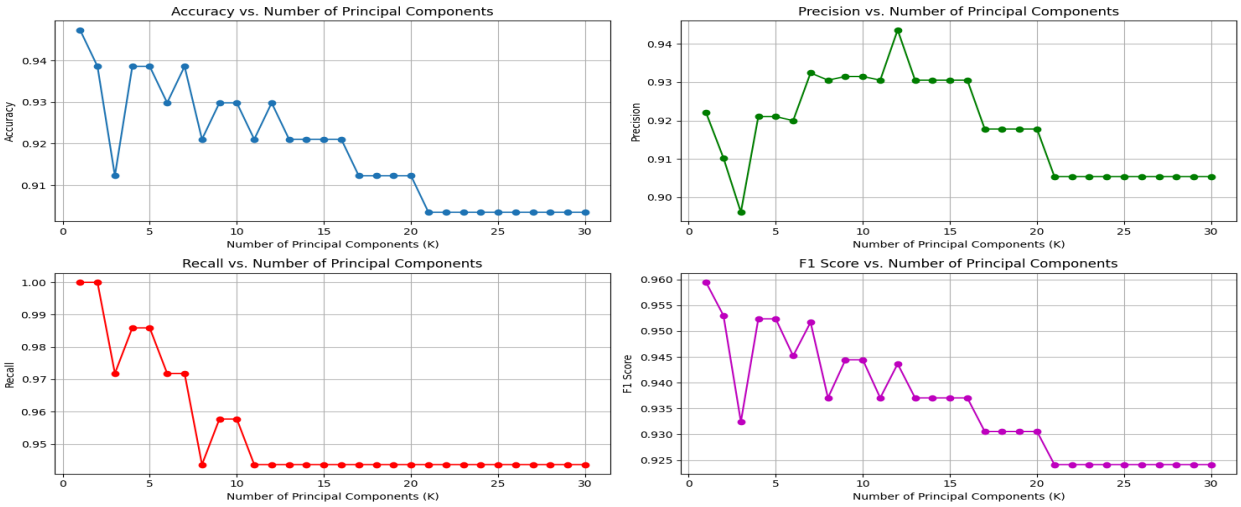
Cancer Dataset Naive Bayes Metrics:
Accuracy: 0.96
Precision: 0.96
Recall: 0.99
F1 Score: 0.97



Comparison with Logistic Regression (Without Penalty):
Accuracy Difference: 0.01
Precision Difference: 0.01
Recall Difference: 0.00
F1 Score Difference: 0.01
Naive Bayes also achieved near-perfect performance, closely matching logistic regression. The slight drop in accuracy and precision indicates that logistic regression may be slightly better at capturing linear boundaries in this dataset

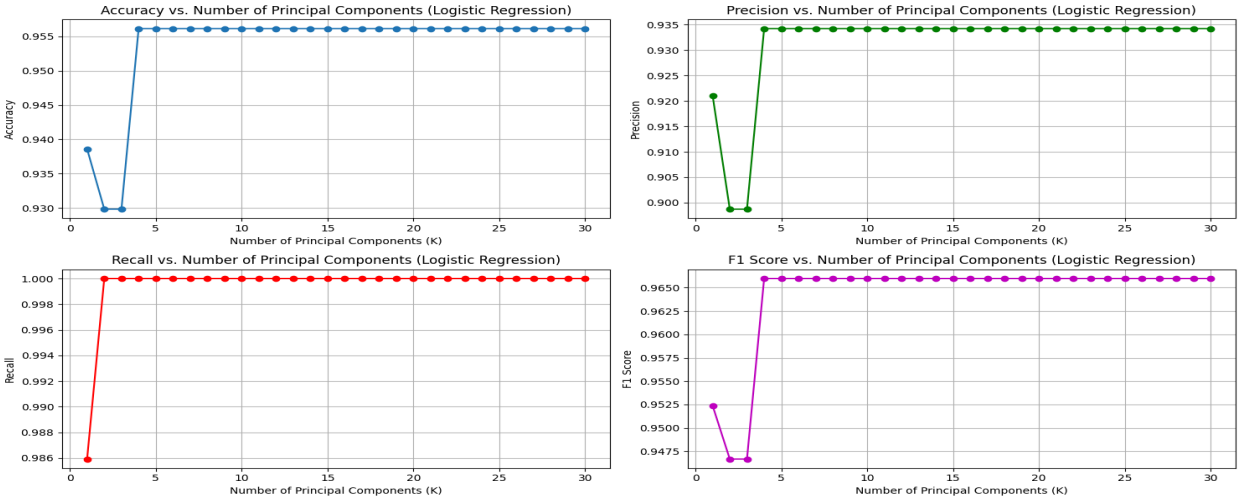
4. Use the cancer dataset to build a logistic regression model to classify the type of cancer (Malignant vs. benign). Use the PCA feature extraction for your training. Perform N number of independent training (N=1, …, K). Identify the optimum number of K, principal components that achieve the highest classification accuracy. Plot your classification accuracy, precision, recall, and F1 score over a different number of Ks. Explain and elaborate on your results and compare it against problems 2 and 3. Correction: Replace the logistic regression with Bayes classifier

Name: Abdirahim Ahmed
Student ID: 801338570



This problem highlights the trade-off between dimensionality reduction and model performance. The results likely show a plateau where increasing the number of components improves performance up to a point. Naive Bayes might perform slightly worse with fewer components due to loss of information.

5.Can you repeat problem 4? This time, replace the Bayes classifier with logistic regression. Report your results (classification accuracy, precision, recall and F1 score). Compare your results against problems 2, 3 and 4.



The performance of logistic regression with PCA is expected to align closely with its performance in Problem 2. However, PCA might introduce slight drops in precision or recall if too few components are selected. The plots showing metrics over different K values provide insights into how well logistic regression performs with reduced dimensions