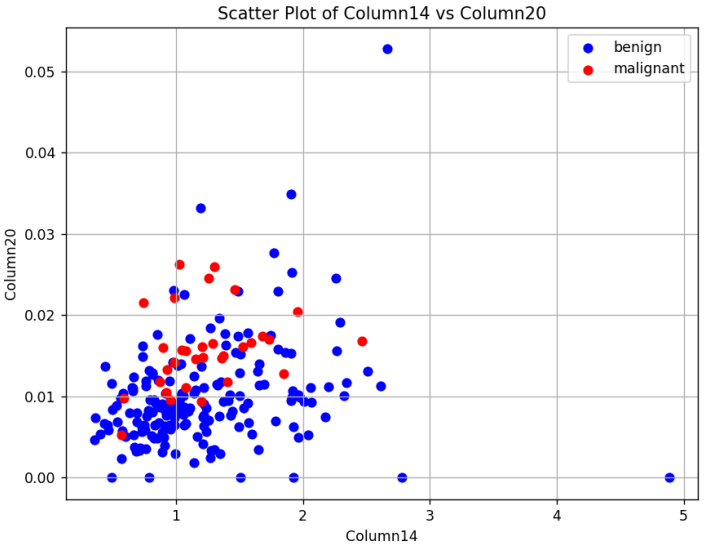
Mahmoud Mohamed Hassan : 20045532

Data Classification Report

Task1:

- the Training and testing dataset is read from the CSV file and split into X train Y Train , X test Y test .
- both datasets are normalized by standard scaler and saved in two variables.
- Two features from the training dataset is plotted on scatter plot (column 14 vs column 20)
- Both datasets are reduced by PCA and saved in two variables.
- PCA analysis is done on the training set by plotting scree plot & projection of Training set (PC1 Vs PC2)
- After analyzing the scree plot number of components = 5 is chosen to reduce the training and testing the dataset .



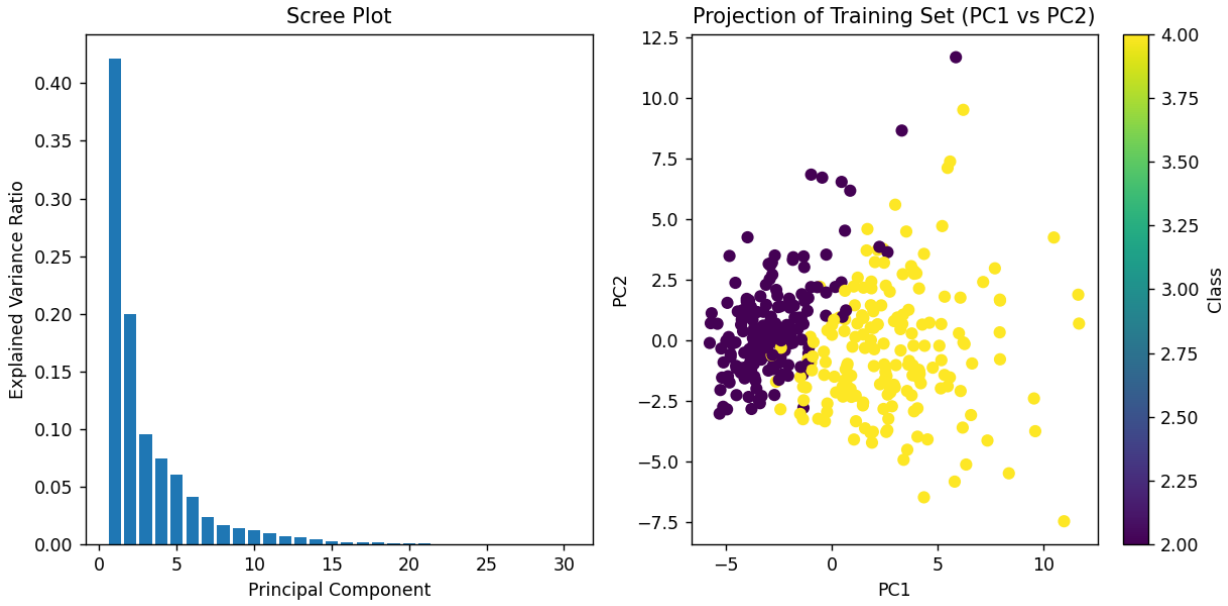
One Scatterplot : the plot is between feature1 (column 14), feature2 (column20) the scatterplot shows that there are overlapping between classes where the points representing different classes overlap significantly is suggests that more complex decision boundaries are needed , there are class in balance where class benign dominates the scatterplots while malignant class is sparse it indicates class in balance , there are outliers identified in benign class

Standard Scaler: Standard scaler is used from SKLearn Library to normalize both train and test sets , one standard scaler object is initialized and given train set to fit on the data and transform it, and given test set to transform it only , the normalization is done on each set by subtracting the mean and divide by the standard deviation this process ensures that the distribution of each feature has a mean nearly equal 0 , And the standard deviation nearly = 1

The Mean of the first Feature after Normalization = -0.3814457765983103

The Standard deviation of the first Feature after Normalization = 0.8381524806309703

PCA analysis:



Scree plot: we making decision on how many components or factors to retain , in the scree plot the elbow or break point is at number of components equal 6 so the number of components to retain is 5

Projection of training set (PC1 vs PC2):

PC1: weight combination that results in the largest sample variation

PC2: weight combination that result in the second largest variation

Plotting PC1 vs PC2 Results Greatest spread of Data which will better enable pattern recognition.

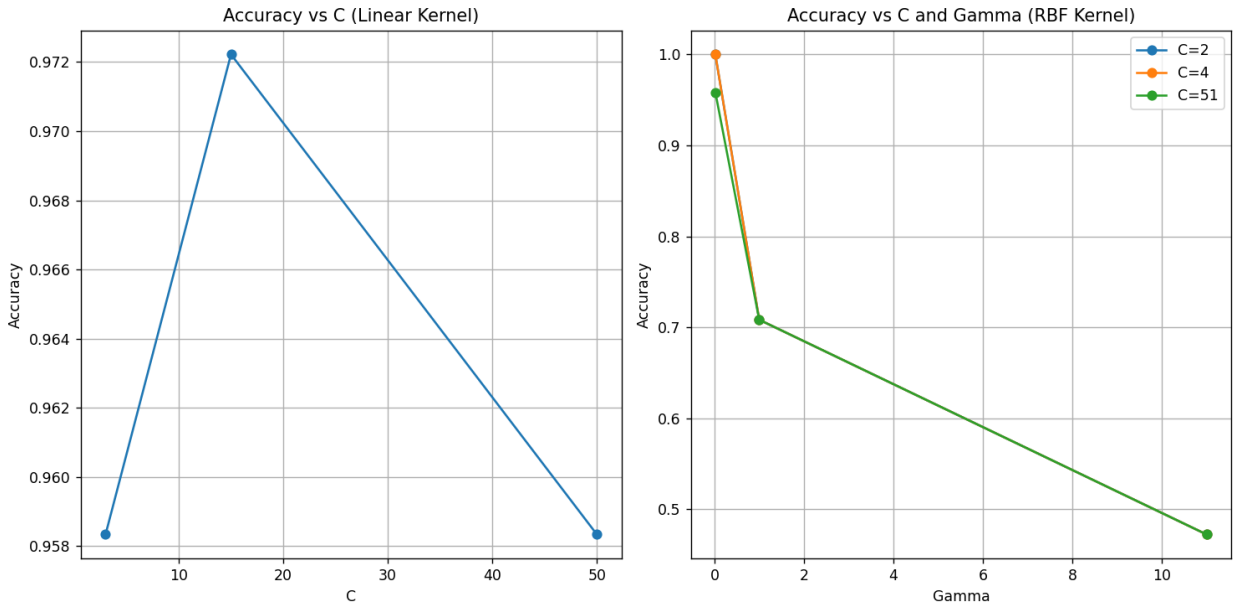
Task 2:

The training data is split by train_test_split from SKLearn into Training set (II), and validation set by Ratio 80 : 20 , the two sets are normalized by standard scaler method from SKLearn two objects where created one for Training set and other for validation set by subtracting the mean and divide by standard deviation .

- Training Feature 1 Mean after normalization = 4.949744318120489e-16
- Training Standard deviation after normalization = 1.0
- Validation Feature 1 mean after normalization = 2.0816681711721685e-16
- Validation standard deviation after normalization = 1.0
- the mean value and the standard deviation of each feature in the normalized training set are nearly equal to the mean and standard deviation of the corresponding feature in the normalized validation set.

Task 3:

- Two SVM models are created
- the first model uses kernel (linear) with three different C values [3 , 15 , 50]
- The second model uses Kernel (RBF) with three different C values [2, 4, 51], three different gamma values [0.02 , 1, 11]
- The two model is trained on training set (II) and tested on validation set



The accuracies got by the Linear Kernel in SVM :

C	Accuracy
3	96%
15	97%
50	96%

The accuracies got by the RBF Kernel in SVM :

C	Gamma	Accuracy
2	0.02	100%
2	1	70%
2	11	47%
4	0.02	100%
4	1	70%
4	11	47%
51	0.02	96%
51	1	70%
51	11	47%

Task 4:

- PCA algorithm is done on The Training set and the testing set to reduce the dimensionality of the data
- number of components chose after analyzing the scree plot is equal 5
- PCA From SKLearn is used to do this task as following : PCA object is initialized and fitted on the training set then used to transform training and testing set
- Standard scalar object is used from SKLearn to normalized data
- 2 SVM model is initialized
- First model is trained on normalized train set 1 and tested on normalized test set
- Second model is trained on reduced PCA and normalized Train set 1 and tested on reduced PCA and normalized Test set

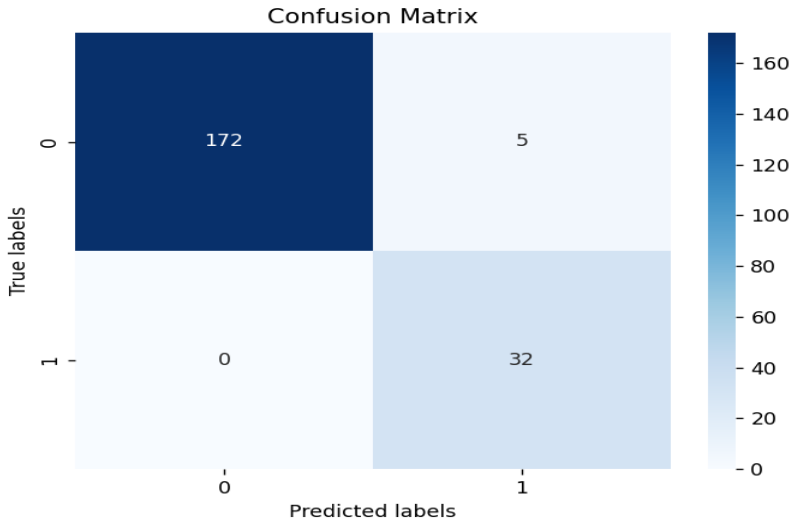
The hyper parameters selected for training the SVM on training set (I) is :

- Kernel: RBF
- C = 4
- Gamma = 0.02
- For reduced data by PCA: in the scree plot the elbow or break point is at number of components equal 6 so the number of components to retain is 5

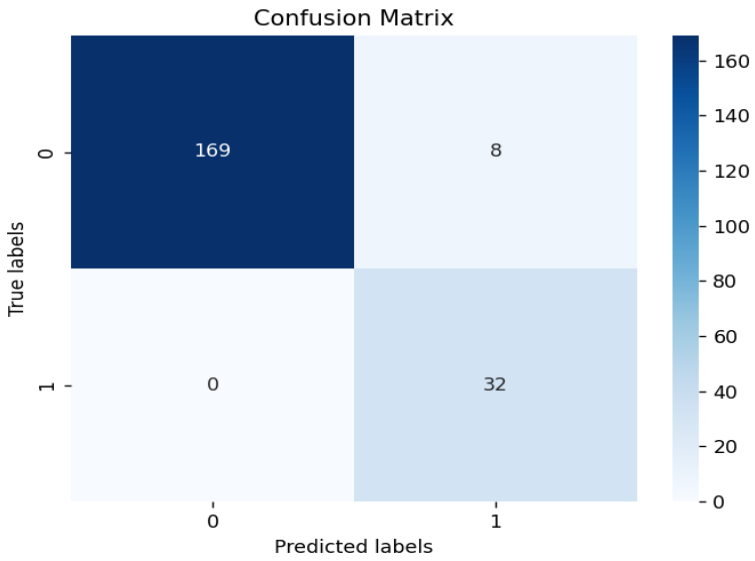
Accuracy got by normalized data only = 98%

Accuracy got by reduced Data by PCA and normalized = 96%

Confusion matrix for normalized data :



Confusion Matrix for PCA transformed data:



Dataset	C	Gamma	Accuracy
Train Set(II)&Validation set	4	0.02	100%
Normalized Train set(I)& Normalized test set	4	0.02	98%
Reduced PCA and normalized Train set (I)& Reduced PCA and normalized Test set	4	0.02	96%

Conclusion :

Normalized Train set(I)& Normalized test set got higher accuracy than Reduced PCA and normalized Train set (I)& Reduced PCA and normalized Test set .