General information about dataset:

- Name: Plant Pathology 2020 - FGVC7 Dataset
- Number of classes: 2
- Class labels: Healthy “ 1 ” , Diseased “ 0 ”

Class Distribution:

- Healthy samples: 516
- Disease samples: 1305

Dataset Size:

- Total samples: 1821
- Train set:1456
- Test set: 365

Feature Extraction

In the pursuit of identifying essential features for the classification task, various feature extraction methods were explored. Through experimentation, the Histogram of Oriented Gradients (HOG) emerged as the most effective feature extractor. HOG is employed due to its capability to describe important features relevant to the plant pathology problem .

block_size: Tuple defining the size of a block in pixels. The image is divided into blocks, and the HOG descriptor is computed within each block. Here, it's set to (16, 16) in our algorithm.

block_stride: Tuple specifying the stride between consecutive blocks. It defines the step size when moving the block window. it's set to (8, 8).

cell_size: Tuple determining the size of a cell in pixels. A cell is a small spatial region within a block, and the HOG descriptor is computed for each cell. Here, it's (8, 8).

nbins: The number of bins in the histogram of gradients. It represents the number of orientation bins in the histogram. it's set to 9, meaning the histogram is divided into 9 bins.

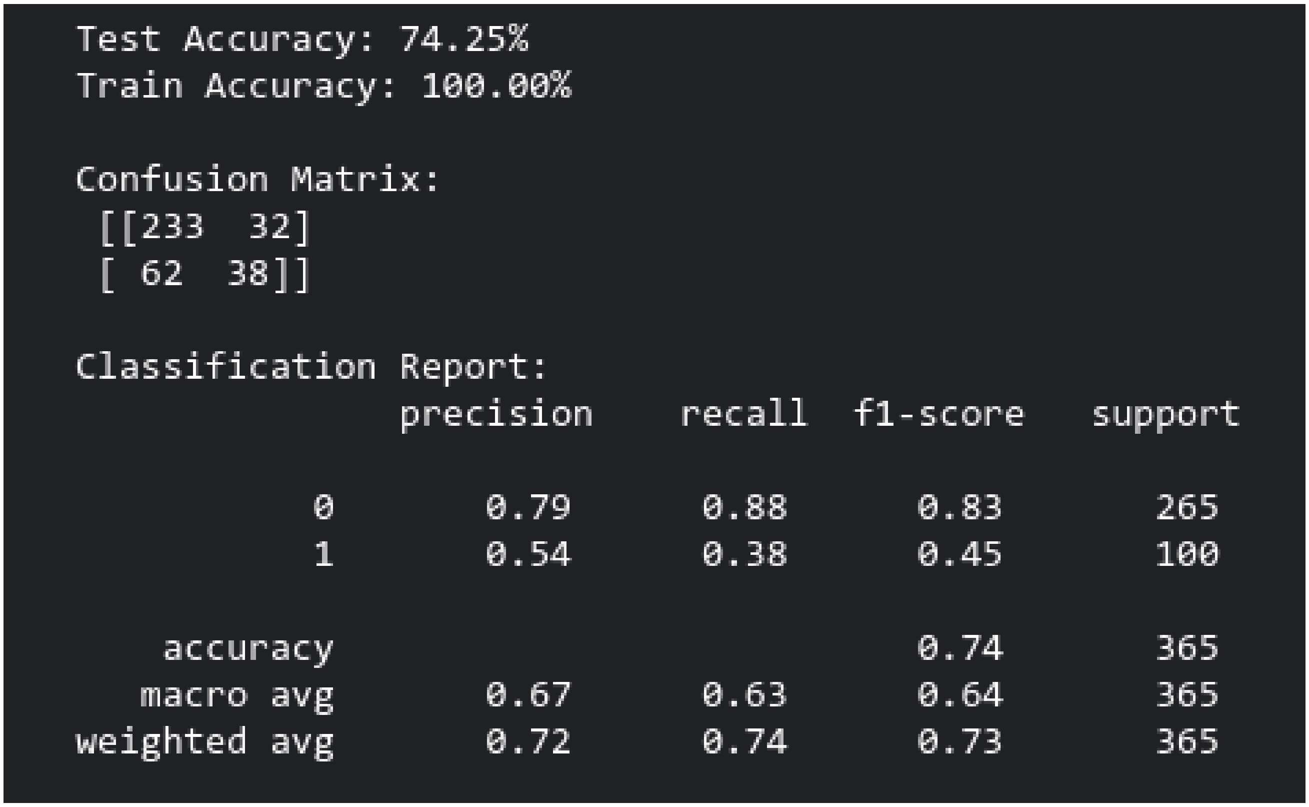
The number of features per image: 13872

First classification Using Logistic Regression:

Modified Hyperparameters:

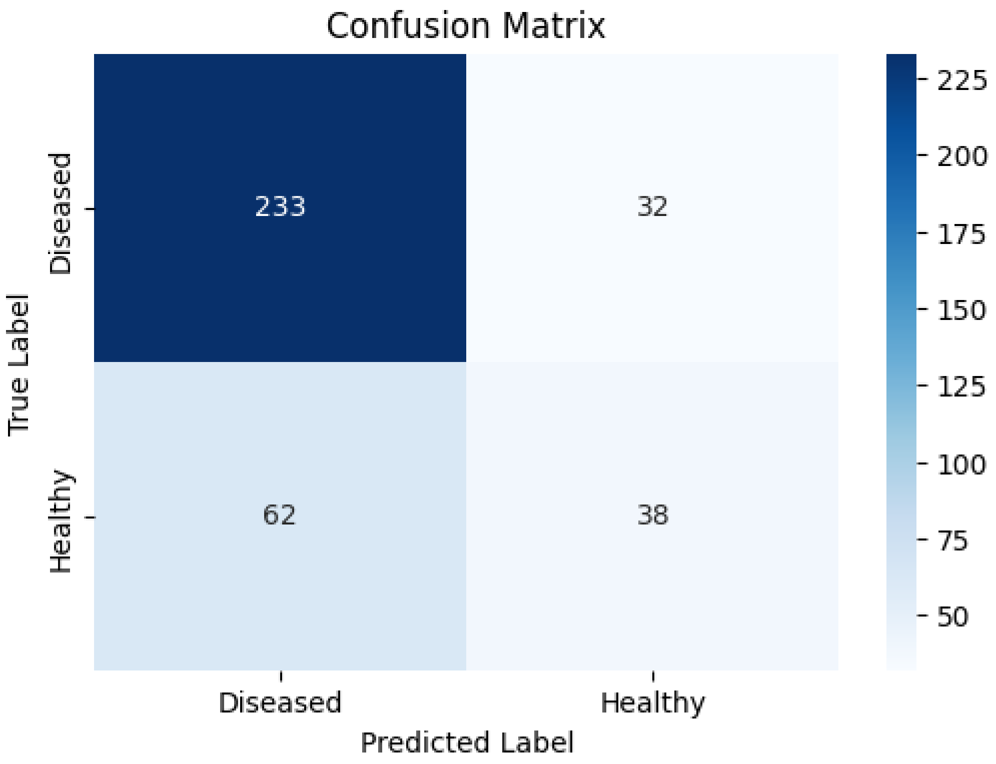
- Penalty: l2 regularization (Ridge)
- C: 5 (Regularization strength Inverse)
- max_iter: 2000

Accuracy & Classification report of Model:

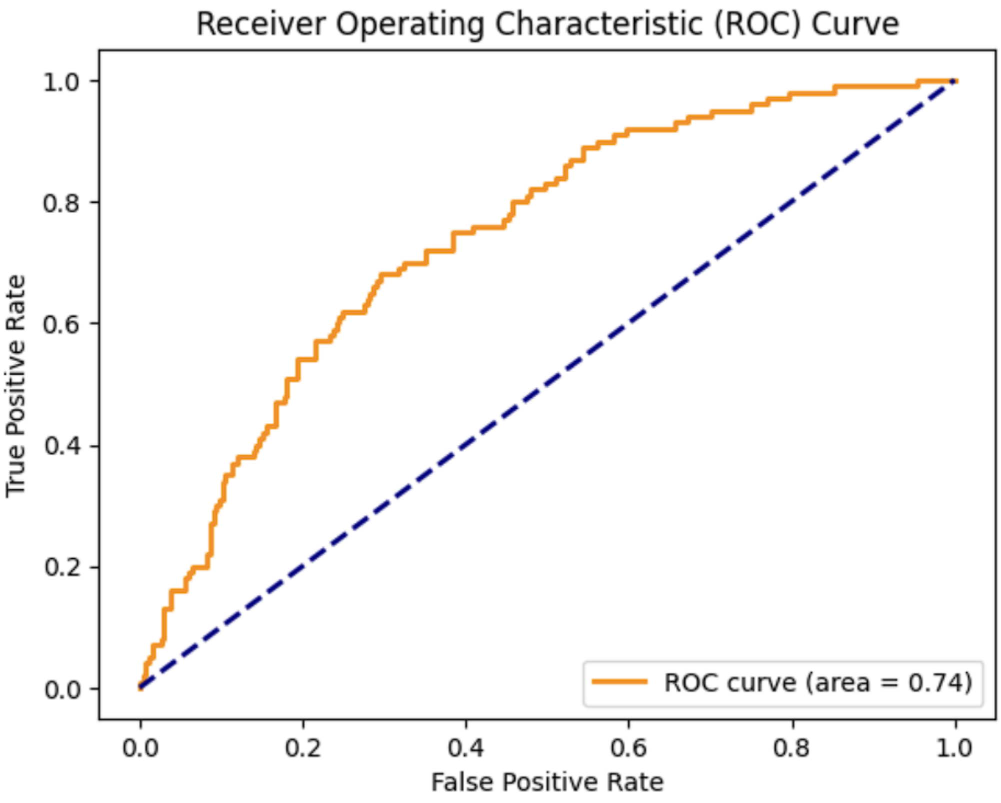


```
Test Accuracy: 74.25%
Train Accuracy: 100.00%

Confusion Matrix:
 [[233  32]
 [ 62  38]]

Classification Report:
              precision    recall  f1-score   support

           0       0.79      0.88      0.83       265
           1       0.54      0.38      0.45       100

    accuracy                           0.74       365
   macro avg       0.67      0.63      0.64       365
weighted avg       0.72      0.74      0.73       365
```

Confusion matrix:



ROC Curve:

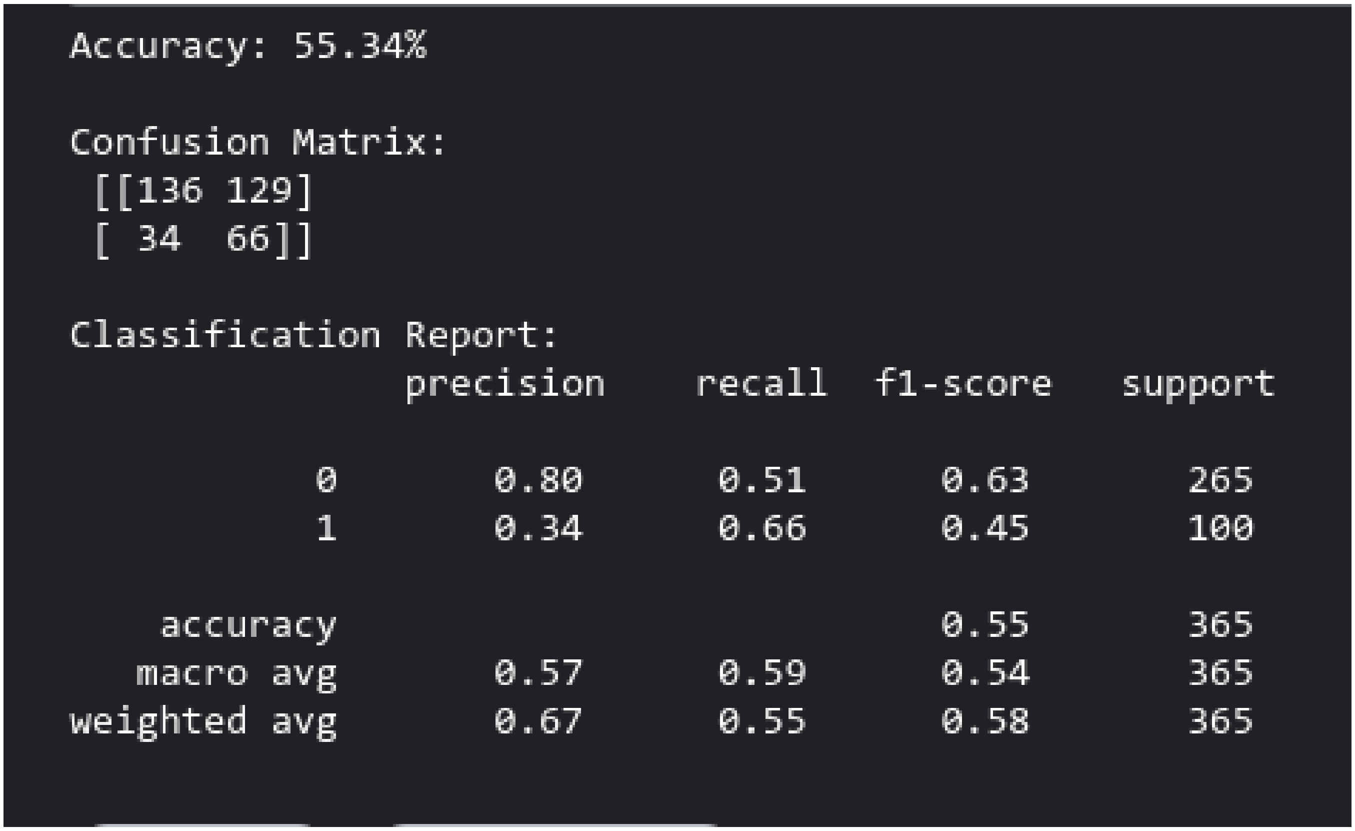


Second classification Using KMeans:

Modified Hyperparameters:

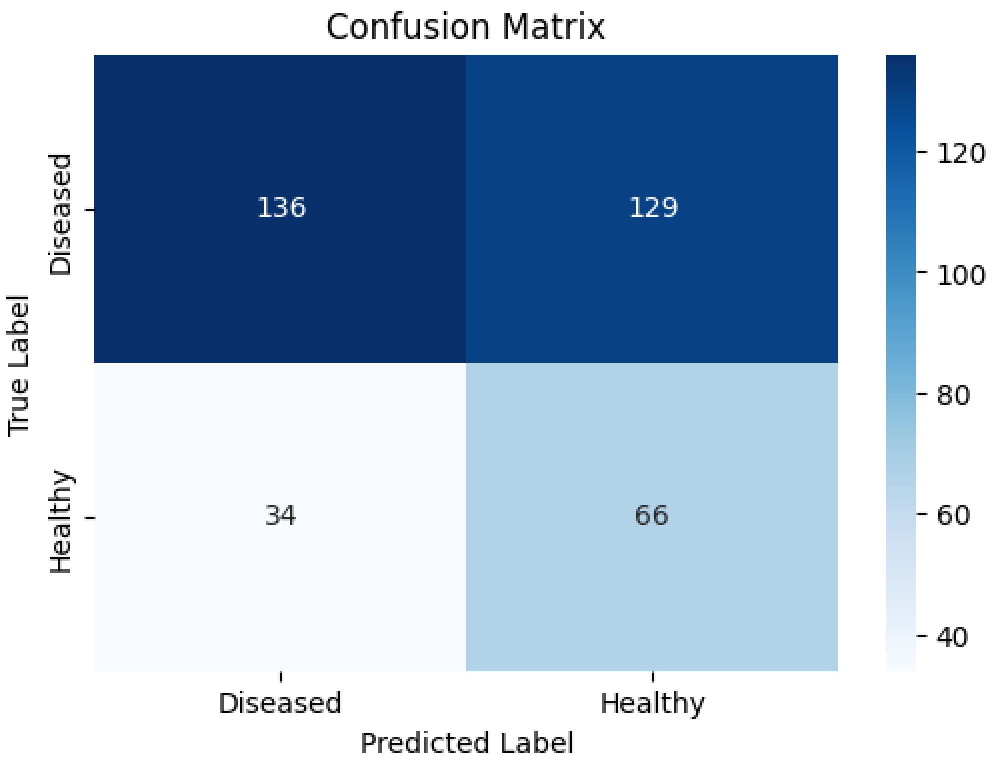
- n_clusters: 2 (Number Of Clusters)
- n_init: 10 (Number of times the KMeans algorithm will be run)

Accuracy & Classification report of Model:

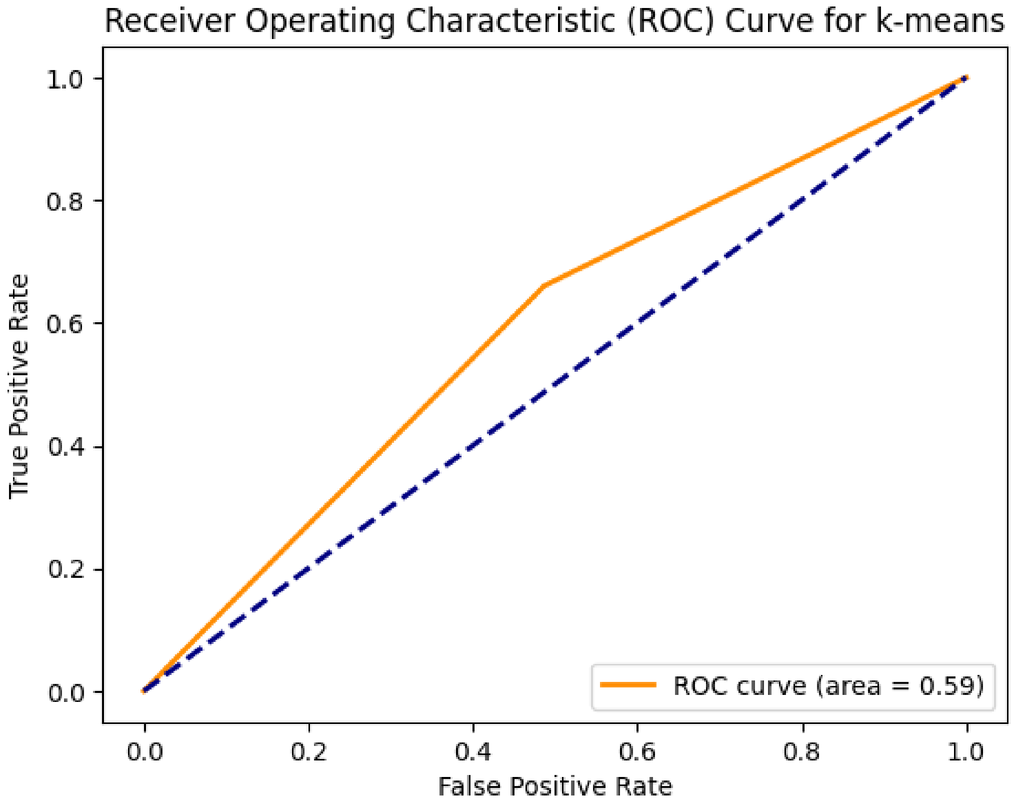


```
Accuracy: 55.34%

Confusion Matrix:
 [[136 129]
 [ 34  66]]

Classification Report:
               precision    recall  f1-score   support

           0       0.80      0.51      0.63       265
           1       0.34      0.66      0.45       100

    accuracy                           0.55       365
   macro avg       0.57      0.59      0.54       365
weighted avg       0.67      0.55      0.58       365
```

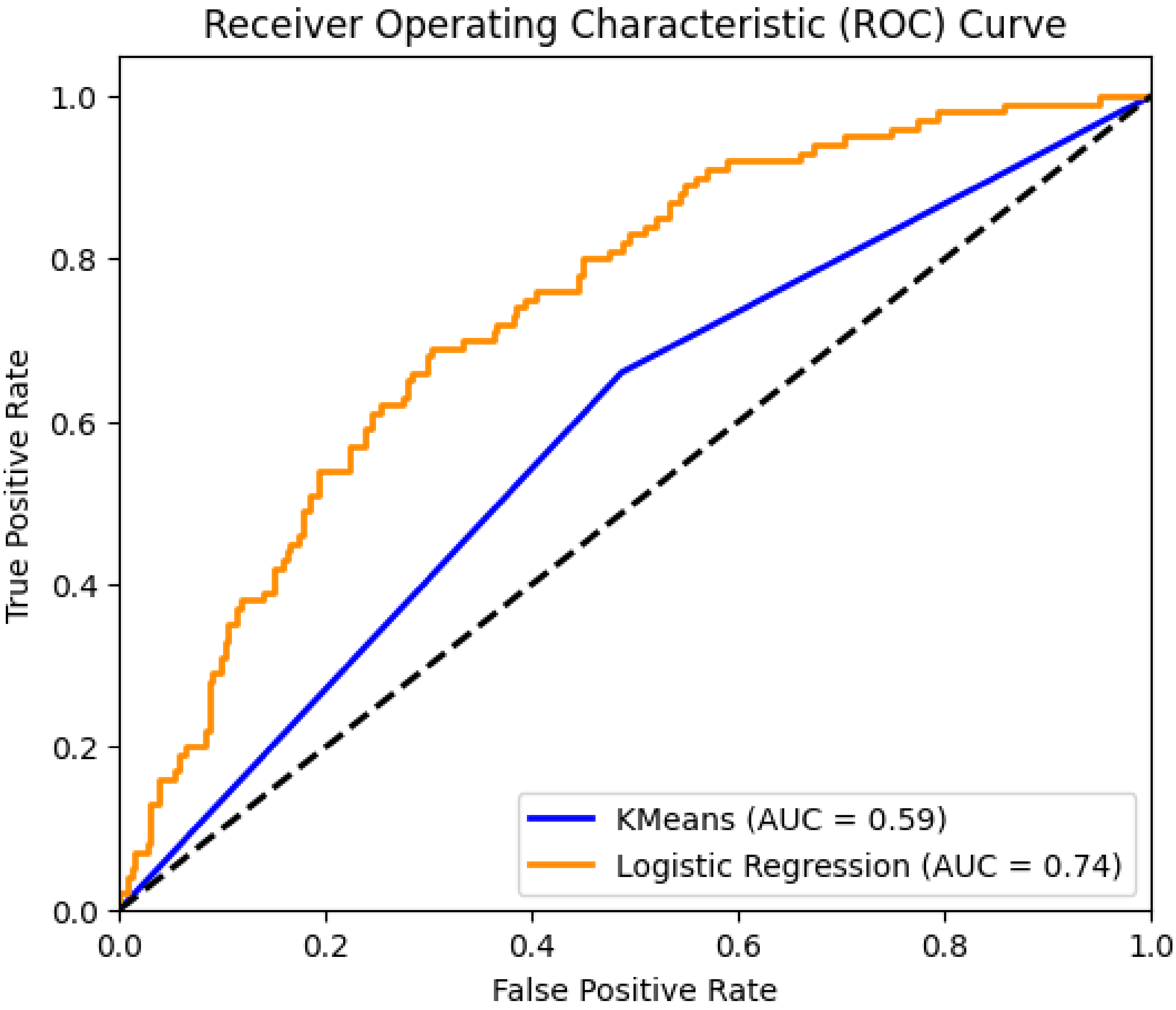
Confusion matrix:



ROC Curve:



The Comparing ROC Curve Between Logistic Regression and KMeans:



The curve indicates that the Logistic Regression model is significantly better than KMeans for our problem.