Assignment 2

COMP - 551

Student ID: 260880458

Mohd Safwan Ahmad

Prob.1

See data files DS1test, DS1val, DS1train for the data set

Prob.2

a.

```
w [ 14.31574432  -8.30249007  -5.35750683  -3.07467431  -9.42293241
-4.6212408   16.31258299 -23.83168504 -28.69836891   9.16355032
-12.98307682 -12.02469945  15.32720518  12.6052344   -5.67820318
13.10838037  28.88467948  -6.8008162   -0.32402186  -5.01550096]

w0 27.141267048146638
```

b.

```
Accuracy = 0.96
Precision = 0.9532019704433498
Recall = 0.9675
F Measure = 0.9602977667493796
```

Prob.3

a.

	F-Measure
GDA	0.9602977667493796
KNN	
K=1	0.5541561712846349
K=2	0.6219151036525172
K=3	0.5358024691358024
K=4	0.6047966631908237
K=5	0.5608856088560885
K=6	0.5608856088560885
K=7	0.5503685503685504
K=8	0.6065217391304347
K=9	0.5724725943970768
K=10	0.6102449888641426

GDA Performs better than the KNN for this example.

K implies the number of nearest k points to x (data point in test set) in the training set based on which we classify the test data point.

K = 2 works best in this case – Because this means that the algorithm is classifying based on the closest 2 points for this algorithm and its working best for this example.

b.

```
for k value = 2
Accuracy = 0.52125
Precision = 0.5138662316476346
Recall = 0.7875
F Measure = 0.6219151036525172
```

Prob.4

See data files DS2test, DS2val, DS2train for the data set

Prob.5

1.a.

```
Accuracy = 0.49875
Precision 0.4986737400530504
Recall = 0.47
F Measure = 0.4839124839124839
```

1.b.

```
w [-0.00966259 -0.02559834 -0.07705795 -0.00013703  0.03761088 -0.11042
393
0.00399542 -0.01742475  0.09723131  0.02207119 -0.00945184  0.05906434
0.04334472  0.03385311 -0.01117373  0.04825551 -0.04702542 -0.0178343
0.02450182  0.01071567]

w0 -0.055699759777591396
```

	F-Measure
GDA	0.4839124839124839
KNN	
K=1	0.5263157894736842
K=2	0.5737538148524923
K=3	0.5271122320302648
K=4	0.5932203389830508
K=5	0.527363184079602

2.

KNN is performs better than GDA for this example.

K = 4 works best in this case – Because this means that the algorithm is classifying based on the closest 4 points for this algorithm and its working best for this example.

3.

```
for k value = 4
Accuracy = 0.52
Precision = 0.5147058823529411
Recall = 0.7
F Measure = 0.5932203389830508
```

Prob.6

For DS1, the GDA performed way better, generally, and better than the KNN. That is because GDA assumes data coming from a Gaussian distribution and that was exactly how the data was generated.

However, for DS2 though the data was still coming from Gaussian distribution, but it was a mix of 3 Gaussians with different means – thus it was not ideal for GDA hence not as good a performance.

Since KNN doesn't assume anything about the data, it performed almost similar for both DS1 and DS2