Assignment Report: Binary Classifiers

- **Objective:**

Observe and compare the efficiency and attributes of the below binary classifiers:

- ❏ Half Space
- ❏ Logistic Regression (using inbuilt function)
- ❏ SVM classifier (using a linear kernel)
- ❏ SVM classifier (using a Polynomial kernel and a Gaussian kernel)
- ❏ Logistic Regression using the SGD procedure.

- **Procedure:**

1. Choose two datasets, one having linearly separable examples and second having linearly non-separable examples.
2. Read the dataset from the file to create DataFrame in python.
3. Perform Feature standardization to all the features in the dataset.
4. Split the complete dataset into a training set and test set with 70:30, 80:20 and 90:10 ratio and repeat the procedure for these different splits.
5. Taking the training set as input, train the model and draw the prediction for the test set.
6. Compare the prediction with the actual labels given to the test set and compute the efficiency.
7. Repeat the procedure for the given different classifiers and for the svm classifier also vary the regularisation parameter.

- **Observations:**

The efficiency of the model for ***linearly non-separable*** data with varying splits, classifiers and regularisation parameter is given in the table below-

Classifier/Splits Ratio		**Efficiency(%)**		
		70:30	**80:20**	**90:10**
Half Space using perceptron		95.87	95.27	95.65
Logistic Regression		97.82	97.82	97.83
Logistic Regression using SGD		97.82	97.82	95.65
SVM	Regularisation Parameter(C)			
Linear Kernel	C = 0.3	98.54	98.18	97.83
	C = 0.7	98.79	98.55	98.55
	C = 0.9	98.3	98.91	98.55
Polynomial Kernel (degree = 3)	C = 0.3	97.33	98.18	99.28
	C = 0.7	99.03	97.45	98.55
	C = 0.9	99.27	98.91	97.83
Gaussian kernel	C = 0.3	99.03	99.27	100.0
	C = 0.7	100.0	100.0	100.0
	C = 0.9	100.0	100.0	100.0

The Number of Iterations required for the convergence of Logistic Regression for different test set splits is given in the table below-

Classifier		**No. of iterations**		
	Split Ratio	**70:30**	**80:20**	**90:10**
Logistic Regression		16	15	16
Logistic Regression using SGD		23	20	20

The number of support vectors obtained with varying splits and regularisation parameter is given in the table below-

SVM With Kernel Function	**No. Of Support Vectors**			
	Split Ratio/Regularisation Parameter(C)	**C= 0.3**	**C= 0.7**	**C= 0.9**
Linear Kernel	70:30	91	62	62
	80:20	98	72	70
	90:10	102	80	72
Polynomial Kernel	70:30	356	280	263
	80:20	400	302	281
	90:10	433	334	303
gaussian Kernel	70:30	163	104	91
	80:20	175	110	99
	90:10	187	117	104

The above computations are made on a dataset having **1372** examples and **5** features.

The efficiency of the model for ***linearly separable*** data with varying splits, classifiers and regularisation parameter is given in the table below-

Splits Ratio		**70:30**	**80:20**	**90:10**
Half Space using perceptron		100.0	100.0	100.0
Logistic Regression		100.0	100.0	100.0
Logistic Regression using SGD		100.0	100.0	100.0
SVM	Regularisation Parameter(C)			
Linear Kernel	C = 0.3	100.0	100.0	100.0
	C = 0.7	100.0	100.0	100.0
	C = 0.9	100.0	100.0	100.0
Polynomial Kernel (degree = 3)	C = 0.3	100.0	100.0	100.0
	C = 0.7	100.0	100.0	100.0
	C = 0.9	100.0	100.0	100.0
Gaussian kernel	C = 0.3	100.0	100.0	100.0
	C = 0.7	100.0	100.0	100.0
	C = 0.9	100.0	100.0	100.0

The Number of Iterations required for the convergence of Logistic Regression for different test set splits is given in the table below-

Classifier		**No. of iterations**		
	Split Ratio	**70:30**	**80:20**	**90:10**
Logistic Regression		11	12	11
Logistic Regression using SGD		19	20	20

The number of support vectors obtained with varying splits and regularisation parameter is given in the table below-

SVM With Kernel Function	**No. Of Support Vectors**			
	Split Ratio/Regularisation Parameter(C)	**C= 0.3**	**C= 0.7**	**C= 0.9**
Linear Kernel	70:30	4	3	4
	80:20	4	4	4
	90:10	4	3	4
Polynomial Kernel	70:30	23	15	15
	80:20	25	17	15
	90:10	25	16	14
gaussian Kernel	70:30	13	11	14
	80:20	12	12	12
	90:10	13	12	12

The above computations are made on a dataset having **100** examples and **5** features.

- ## Conclusions:

1. In the smooth scenario, where the data is linearly separable the efficiency is 100% for all the classifiers. In this case, Half Space using perceptron is a better approach because of its simplicity.
2. Whenever the data is linearly non-separable, Soft SVM with gaussian kernel function gives the best efficiency and good generalization for high values of regularisation parameter(C = 0.9). While, half space classifiers performs the worst.
3. The number of support vectors as the output of the SVM is inversely proportional to the regularisation parameter(C).
4. The number of support vectors as the output of the SVM is highest with the gaussian kernel function and lowest for the linear kernel function.