**I.CHOOSING DATA SET**

**Data set used:**

<http://archive.ics.uci.edu/ml/machine-learning-databases/00379/PhishingData.arff>

**Dataset description:**

Task – Classification

Attribute Type - Numerical

Data Type – Multivariant

Number of instances = 1353

Numbed of attributes = 9

Missing or N/A values = 0

**II.PRE-PROCESSSING DATA SET**

The data set has been split for training & test in 75% ratio.

The pre-processing of dataset has been done using utility class **StandardScaler** that implements the Transformer API to compute the mean and standard deviation on a training set so as to be able to later reapply the same transformation on the testing set.

**III. Finding best classifier parameters**

By running the classifiers with multiple parameters, below is the list of params best for each classifier:

Please find the Analysis.xl file the experiments with different values.

**Decision Tree Classifier:**

criterion='entropy', max\_depth=10, max\_features=None, max\_leaf\_nodes=60, min\_samples\_leaf=5, min\_samples\_split=2, min\_weight\_fraction\_leaf=0.0, presort=False, random\_state=1000, splitter='best'

**Perceptron:**

penalty='l2', alpha=0.0001, fit\_intercept=True, shuffle=False, verbose=0, eta0=1.0, random\_state=1, class\_weight=None, warm\_start=True

**Neural network:**

**solver='adam'**, alpha=1e-5,hidden\_layer\_sizes=(30,30,30),max\_iter=200,random\_state=1, **activation='relu'**,learning\_rate='constant', **learning\_rate\_init=0.001, shuffle=True**, momentum=0.9, **early\_stopping=False**

**SVM:**

**C=2.0, kernel='poly', degree=3,** gamma='auto', coef0=0.0, shrinking=True, probability=False, tol=0.001, cache\_size=200, class\_weight=None, verbose=False, **max\_iter=1000, decision\_function\_shape='ovo', random\_state=100**

**Naive Bayes:**

GaussianNB(priors=None)

**Logistic Regression:**

**penalty='l2', dual=False, tol=0.001, C=1.0, fit\_intercept=True**, intercept\_scaling=1, class\_weight=None, random\_state=None, **solver='lbfgs', max\_iter=100,** multi\_class='ovr', verbose=0, warm\_start=False, n\_jobs=1

**K Nearest Neighbors:**

algorithm='auto', leaf\_size=30, metric='minkowski',metric\_params=None, n\_jobs=1, **n\_neighbors=5,** p=2, weights='uniform'

**Bagging:**

**base\_estimator=None, n\_estimators=100**, max\_samples=1.0, max\_features=1.0, **bootstrap=True, bootstrap\_features=False**, oob\_score=False, warm\_start=False, n\_jobs=1, **random\_state=100**, verbose=0

**Random Forest Classifier:**

**n\_estimators=25, criterion='entropy', max\_depth=None,** min\_samples\_split=2, min\_samples\_leaf=1, min\_weight\_fraction\_leaf=0.0, **max\_features='sqrt'**, max\_leaf\_nodes=None, **bootstrap=True**, oob\_score=False, n\_jobs=1, **random\_state=None**, verbose=0, warm\_start=False, class\_weight=None

**Ada Boost Classifier:**

**base\_estimator=None, n\_estimators=30, learning\_rate=1.0, algorithm='SAMME.R', random\_state=50**

**Gradient Boosting Classifier:**

loss='deviance', learning\_rate=0.1, n\_estimators=60, subsample=0.8, criterion='friedman\_mse', min\_samples\_split=100, min\_samples\_leaf=1, min\_weight\_fraction\_leaf=0.0, max\_depth=9,init=None, random\_state=None, max\_features=None, verbose=0, max\_leaf\_nodes=None, warm\_start=False, presort='auto'

**Note:**

Deep learning is not yet implemented in Scikit Learn.

**IV. Testing all classifiers together**

**Comparing Average Accuracy of multiple classifiers**

DT :

88.841503268

Perceptron :

79.6688453159

MLP :

83.667211329

SVM:

87.6530501089

NB:

82.0392156863

LR:

83.2216775599

GradBoost :

89.9482570806

k-nn :

86.5473856209

Bag :

87.288671024

RandF :

88.6181917211

AdaBoost :

84.3316993464

**V. Report – Confusion Matrix, Precision, Recall, F1-score**

Number of instances = 1353

Numbed of attributes = 9

10 fold cross-validation performed

**Decision Tree Classifier:**

Accuracy of Perceptron Classifier: 88.401416122

***Confusion Matrix:***

[[70 4 5]

[ 1 11 1]

[ 4 0 39]]

Other Metrics:

precision recall f1-score support

-1 0.93 0.89 0.91 79

0 0.73 0.85 0.79 13

1 0.87 0.91 0.89 43

avg / total 0.89 0.89 0.89 135

**Perceptron Classifier:**

Accuracy of Perceptron Classifier: 81.7483660131

Confusion Matrix:

[[67 0 9]

[ 2 0 3]

[ 5 1 48]]

Other Metrics:

precision recall f1-score support

-1 0.91 0.88 0.89 76

0 0.00 0.00 0.00 5

1 0.80 0.89 0.84 54

avg / total 0.83 0.85 0.84 135

**Neural network:**

Accuracy of Neural network: 84.2598039216

Confusion Matrix:

[[60 0 6]

[ 2 0 3]

[ 6 0 58]]

Other Metrics:

precision recall f1-score support

-1 0.88 0.91 0.90 66

0 0.00 0.00 0.00 5

1 0.87 0.91 0.89 64

avg / total 0.84 0.87 0.86 135

**SVM:**

Accuracy of SVM: 87.8039215686

Confusion Matrix:

[[61 1 6]

[ 0 6 3]

[ 6 0 52]]

Other Metrics:

precision recall f1-score support

-1 0.91 0.90 0.90 68

0 0.86 0.67 0.75 9

1 0.85 0.90 0.87 58

avg / total 0.88 0.88 0.88 135

**Naive Bayes:**

Accuracy of Gaussian Naive Bayes: 81.5909586057

Confusion Matrix:

[[61 0 6]

[ 6 2 5]

[ 7 3 45]]

Other Metrics:

precision recall f1-score support

-1 0.82 0.91 0.87 67

0 0.40 0.15 0.22 13

1 0.80 0.82 0.81 55

avg / total 0.78 0.80 0.78 135

**Logistic Regression:**

Accuracy of Logistic Regression: 83.1525054466

Confusion Matrix:

[[71 1 7]

[ 4 0 5]

[ 5 1 41]]

Other Metrics:

precision recall f1-score support

-1 0.89 0.90 0.89 79

0 0.00 0.00 0.00 9

1 0.77 0.87 0.82 47

avg / total 0.79 0.83 0.81 135

**k-Nearest Neighbors:**

Accuracy of k-Nearest Neighbors: 86.3278867102

Confusion Matrix:

[[75 0 3]

[ 0 9 1]

[ 6 2 39]]

Other Metrics:

precision recall f1-score support

-1 0.93 0.96 0.94 78

0 0.82 0.90 0.86 10

1 0.91 0.83 0.87 47

avg / total 0.91 0.91 0.91 135

**Bagging Classifier:**

Accuracy of Bagging Classifier: 88.0283224401

Confusion Matrix:

[[69 0 2]

[ 1 3 7]

[ 4 1 48]]

Other Metrics:

precision recall f1-score support

-1 0.93 0.97 0.95 71

0 0.75 0.27 0.40 11

1 0.84 0.91 0.87 53

avg / total 0.88 0.89 0.88 135

**RandomForestClassifier:**

Accuracy of RandomForestClassifier: 89.5795206972

Confusion Matrix:

[[63 2 6]

[ 0 9 0]

[ 6 0 49]]

Other Metrics:

precision recall f1-score support

-1 0.91 0.89 0.90 71

0 0.82 1.00 0.90 9

1 0.89 0.89 0.89 55

avg / total 0.90 0.90 0.90 135

**AdaBoostClassifier:**

Accuracy of AdaBoostClassifier: 84.2608932462

Confusion Matrix:

[[65 1 2]

[ 5 1 5]

[ 6 0 50]]

Other Metrics:

precision recall f1-score support

-1 0.86 0.96 0.90 68

0 0.50 0.09 0.15 11

1 0.88 0.89 0.88 56

avg / total 0.84 0.86 0.83 135

**GradientBoostingClassifier:**

Accuracy of GradientBoostingClassifier: 89.5762527233

Confusion Matrix:

[[51 1 11]

[ 2 6 2]

[ 4 2 56]]

Other Metrics:

precision recall f1-score support

-1 0.89 0.81 0.85 63

0 0.67 0.60 0.63 10

1 0.81 0.90 0.85 62

avg / total 0.84 0.84 0.84 135