Adv. Topics, Networking CSCI 6515

Moustafa Dafer B00772009
Ms450806@dal.ca

Assignment 1

Metrics and Evaluation

Dataset #instances: 5135

Dimensions: 29

Number of features: 25

Number of classes: 3

Split into:

3594 Training Set (70%)

1541 Testing Set (30%)

First, we binarize the "class" by splitting it into 3 categories: Elk, Deer, Cattle

Using default parameters, we train all of Random Forest Trees, Logic Regression, Naïve Bayes, and Decision Trees.

1) Random Forests model applied on test data:

Scores:

a- ELK: 0.740428293316

b- CATTLE: 0.831278390655

c- DEER: 0.756651524984

Average: 0.776119402985

Mean matrix of Convolution Matrices:

$925.\overline{3}$	243
102	$270.\overline{6}$

2) Random Forests model applied on training data:

Scores:

a- ELK: 0.983583750696

b- CATTLE: 0.986644407346

c- DEER: 0.983583750696

Average: 0.984603969579

Mean matrix of Convolution Matrices:

$23925.\overline{6}$	52
$3.\overline{3}$	1146

3) Random Forests Cross Validation Score:

a- ELK:

[0.73611111 0.73888889 0.73055556 0.725 0.72222222 0.73611111

0.725 0.72905028 0.73463687 0.72346369]

Mean: 0.730103972688

SD: 0.00574041812101

b- CATTLE

[0.84444444 0.84722222 0.84722222 0.83611111 0.82777778 0.825

0.84444444 0.81284916 0.7877095 0.81284916]

Mean: 0.828563004345

SD: 0.0185308319473

c- DEER

[0.73888889 0.74722222 0.78055556 0.72777778 0.74651811 0.75487465

0.74651811 0.74930362 0.78272981 0.76044568]

Mean: 0.753483441659

SD: 0.0163283345232

Average of Means across different types: 0.770716806231

Average of SD across different types: 0.0135331948639

4) Decision Trees model applied on test data:
    a- ELK: 0.685918234912
    b- CATTLE: 0.768981181051
    c- DEER: 0.70214146658

    Average: 0.719013627515

5) Decision Trees convolution matrix mean:

$800.\overline{6}$	$206.\overline{3}$
$226.\overline{6}$	$307.\overline{3}$

6) Decision Trees model applied on training data:
    a- ELK: 1.0
    b- CATTLE: 1.0
    c- DEER:1.0

    Average: 1.0

7) Decision Trees model applied on training data convolution matrix:

2396	0
0	1198

8) Decision Trees Cross Validation Score:
    a- ELK

    [0.65833333 0.65277778 0.66944444 0.64444444 0.66388889 0.62777778

    0.70277778 0.65363128 0.67318436 0.68715084]

    Mean: 0.663341092489

    SD: 0.0202715842335

    b- CATTLE

    [0.8 0.81666667 0.79444444 0.75555556 0.78055556 0.775

    0.80555556 0.74860335 0.7849162 0.77094972]

    Mean: 0.783224705152

    SD: 0.0205349731198

    c- DEER

    [0.65833333 0.7 0.72222222 0.66666667 0.68245125 0.67409471

0.68245125 0.67688022 0.65738162 0.66295265]

Mean: 0.678343392139

SD: 0.0191345789991

Average of Means across different types: 0.70830306326

Average of SD across different types: 0.0199803787841

9) Logic Regression model applied on training data:
    a- ELK: 0.693056456846
    b- CATTLE: 0.813757300454
    c- DEER: 0.733290071382

Average: 0.746701276228

Logic Regression model applied on testing data convolution matrix mean:

$936.\overline{6}$	$299.\overline{6}$
$90.\overline{6}$	214

10) Logic Regression model applied on training data:
    a- ELK: 0.693377851976
    b- CATTLE: 0.803283249861
    c- DEER: 0.722314969393

Average: 0.73965869041

Logic Regression model applied on training data convolution matrix mean:

$2177.\overline{3}$	717
$218.\overline{6}$	481

11) Logic Regression Cross Validation Score:

a- ELK

[0.68888889 0.66388889 0.68888889 0.68333333 0.725 0.66944444

0.70833333 0.70391061 0.68435754 0.67597765]

Mean: 0.689202358783

SD: 0.0176860173343

b- CATTLE

[0.80555556 0.8 0.81388889 0.81111111 0.79722222 0.79166667

0.81666667 0.78212291 0.80167598 0.79608939]

Mean: 0.801599937927

SD: 0.010071021854

c- DEER

[0.71944444 0.74444444 0.73611111 0.70555556 0.70752089 0.71309192

0.7270195 0.72423398 0.72144847 0.72144847]

Mean: 0.722031878675

SD: 0.0114000432507

Average of means across different types: 0.737611391795

Average of SD across different types: 0.013052360813

12) Naïve Bayes model applied on testing data:

a- ELK: 0.550940947437

b- CATTLE: 0.788449059053

c- DEER: 0.350421804023

Average: 0.563270603504

Naïve Bayes convolution matrix mean:

$462.\overline{6}$	$108.\overline{3}$
$564.\overline{6}$	$405.\overline{3}$

13) Naïve Bayes model applied on training data:
    a- ELK: 0.539510294936
    b- CATTLE: 0.792153589316
    c- DEER: 0.353366722315

Average: 0.561676868855

Naïve Bayes applied on training set convolution matrix:

1074	$253.\overline{3}$
1322	$944.\overline{6}$

14) Naïve Bayes Cross Validation Score:
    a- ELK

    [0.55 0.54444444 0.53888889 0.53611111 0.53611111 0.52777778

    0.55 0.53351955 0.54189944 0.54189944]

    Mean: 0.540065176909

    SD: 0.00670661427882

    b- CATTLE

    [0.77777778 0.80555556 0.79722222 0.79444444 0.775 0.79166667

    0.81111111 0.77374302 0.7849162 0.81005587]

    Mean: 0.792149286158

    SD: 0.0133525516596

    c- DEER

    [0.35833333 0.34166667 0.34722222 0.35833333 0.33426184 0.36211699

    0.37047354 0.36211699 0.34261838 0.36490251]

    Mean: 0.354204580625

    SD: 0.0112975032965

    Average of means: 0.562139681231

    Average of SD: 0.0104522230783

T-test results:

Random Forests vs Decision Trees: 1.95666374517e-05

Random Forests vs Logic Regression: 0.00927499133536

Random Forests vs Naïve Bayes: 1.06710942933e-07

Experimenting with n-estimators

N= 10

0.767383487306

N= 20

0.779902335499

N= 50

0.786490884685

N= 100

0.790109488388

T-tests:

RFR100 vs DT

3.6994609051e-08

RFR100 vs LR

3.61720693953e-05

RFR100 vs GNB

1.06066048672e-08

Analysis and assignment answers

a) After splitting the dataset and training the classifiers, we compute the convolution matrix and accuracies, as expected, applying classifiers on the data already trained yields much better results (Figures 1 & 2) than when applying on the testing data; that is due to the fact that the model is already fit to that data; so, in the case of testing data, the classifier is predicting the value of data that is unseen before; we find that random forests performed best amongst the tested classifiers (Figures 1 & 4).
On the other hand, applying the classifier on training data yields almost 100% accuracy (decision trees actually yield 100%), 100% accuracy is not achieved in some classifiers in this case because of pre and post processing, especially for avoiding overfitting.

b) As we have found that Random Forests performed best, we compared it to the other classifiers (Figure 6), and, taking 0.05 as the significance threshold, we find that all comparisons are actually significant; that is logical because each classifier treats the data in a different way and each model is fit differently.

c) N_estimators are the number of trees in the random forest, it basically means that the higher the number of trees, the better the final result, but it also means that more resources are needed with higher number of trees.
As our dataset isn't huge and the difference in requirements between 10,20,50, and 100 n_estimators is negligible in our case, I chose N=100 as the best approach because it yields the best results (Figure 3) and compared the results (Figure 5) as shown in section: "Experimenting with n-estimators".

Result figures

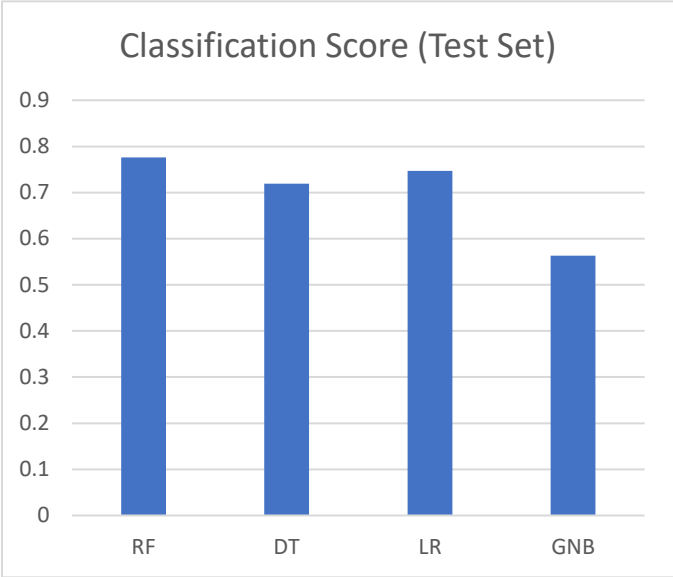

Figure 1. Test Set Classification Scores

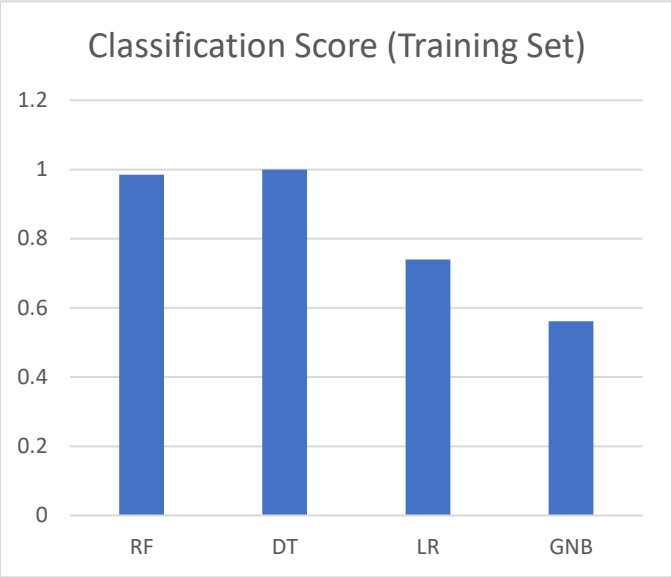

Figure 2. Training Set Classification Scores

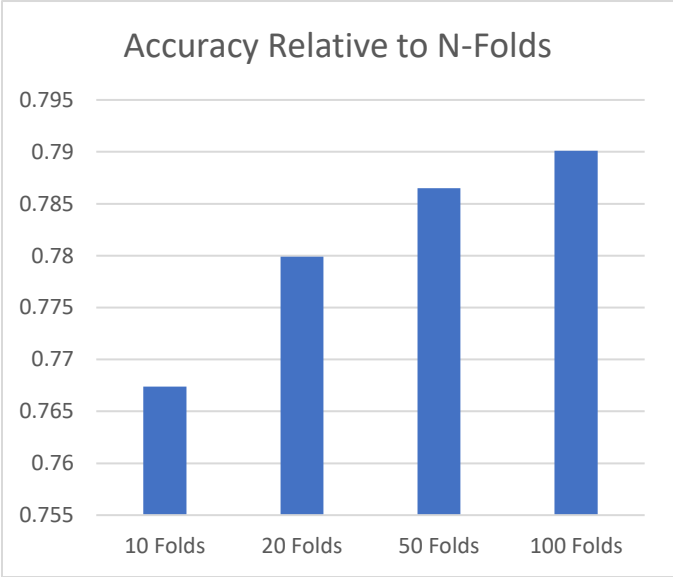

Figure 3. N-Folds Relative Accuracy

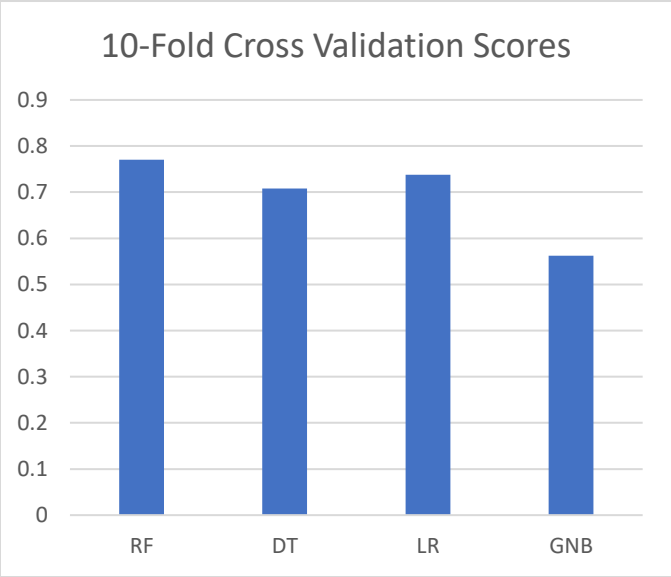

Figure 4. 10-Fold Cross Validation Scores

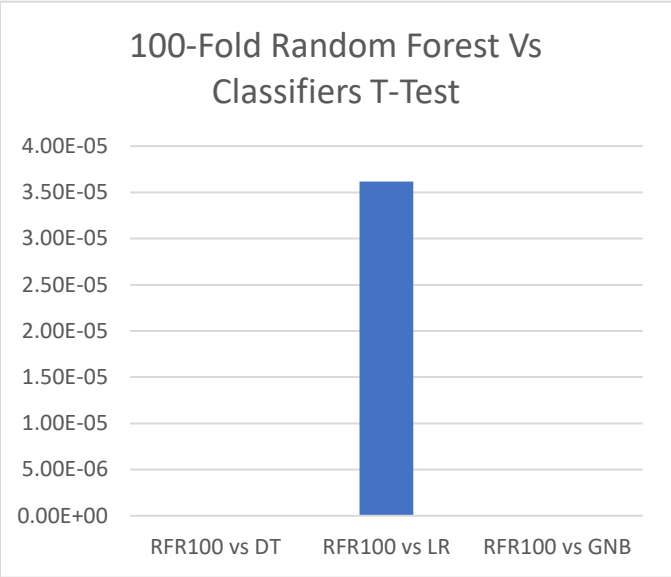

Figure 5. 100-Fold Random Forest Vs Others T-Test

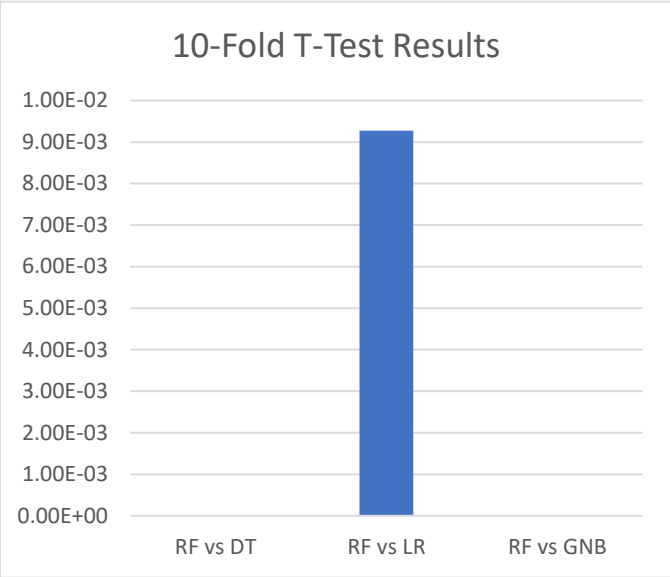

Figure 6. 10-Fold T-Test Results