Assignment 6: Logistic

CIS 435

Section 56

Summer Quarter

School of Continuing Studies

Northwestern University

Daniel Prusinski

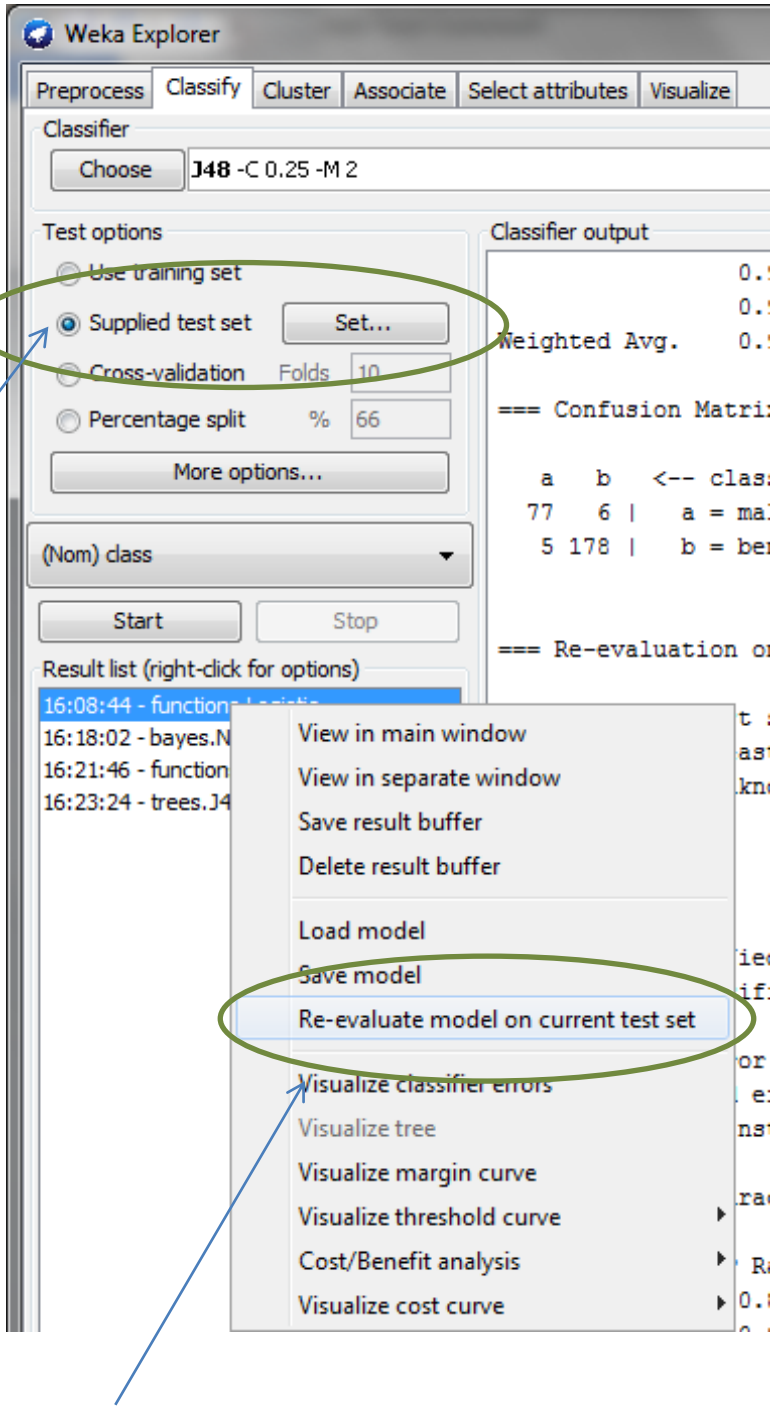
Business Intelligence Data Analyst

Target Corporation

Minneapolis, MN

In Compliance with Master of Science Predictive Analytics

Screen Shots and Output for Initial Comparative Analysis:



This is the option I selected for each algorithm after I loaded the testing data and changed the setting to Supplied test set.

Output:

This is the output generated from each algorithm, the training data is the first output, and the testing data is the second output.

=== Run information ===

Scheme:weka.classifiers.bayes.NaiveBayes
Relation: Breast
Instances: 266
Attributes: 10
clump
ucellsize
ucellshape
magadhesion
sepics
bnuclei
bchromatin
normnucl
mitoses
class
Test mode:10-fold cross-validation

=== Classifier model (full training set) ===

Naive Bayes Classifier

	Class	
Attribute	malignant	benign
	(0.31)	(0.69)
clump		
mean	7.3735	2.847
std. dev.	2.2746	1.682
weight sum	83	183
precision	1	1
ucellsize		
mean	6.3133	1.3552
std. dev.	2.5551	0.9915
weight sum	83	183
precision	1	1

ucellshape

mean	6.3855	1.4918
std. dev.	2.3431	1.1302
weight sum	83	183
precision	1	1

magadhesion

mean	5.5422	1.306
std. dev.	3.152	0.7991
weight sum	83	183
precision	1	1

sepics

mean	5.2861	2.3914
std. dev.	2.3671	0.9631
weight sum	83	183
precision	1.125	1.125

bnuclei

mean	7.631	1.4754
std. dev.	3.0898	1.162
weight sum	83	183
precision	1.125	1.125

bchromatin

mean	5.8193	1.9891
std. dev.	2.066	0.9408
weight sum	83	183
precision	1	1

normnucl

mean	5.9036	1.2951
std. dev.	3.3496	0.9469
weight sum	83	183
precision	1	1

mitoses

mean	2.6747	1.5246
std. dev.	2.4203	0.25
weight sum	83	183
precision	1.5	1.5

Time taken to build model: 0.02 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances	256	96.2406 %
Incorrectly Classified Instances	10	3.7594 %
Kappa statistic	0.9147	
Mean absolute error	0.0375	
Root mean squared error	0.1923	
Relative absolute error	8.7194 %	
Root relative squared error	41.4934 %	
Total Number of Instances	266	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.988	0.049	0.901	0.988	0.943	0.985	malignant
	0.951	0.012	0.994	0.951	0.972	0.989	benign
Weighted Avg.	0.962	0.024	0.965	0.962	0.963	0.987	

=== Confusion Matrix ===

a	b	<-- classified as
82	1	a = malignant
9	174	b = benign

=== Re-evaluation on test set ===

User supplied test set
Relation: Breast
Instances: unknown (yet). Reading incrementally
Attributes: 10

=== Summary ===

Correctly Classified Instances	414	95.612 %
Incorrectly Classified Instances	19	4.388 %
Kappa statistic	0.9065	

Mean absolute error 0.0437
Root mean squared error 0.2043
Total Number of Instances 433

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.968	0.051	0.916	0.968	0.942	0.982	malignant
	0.949	0.032	0.981	0.949	0.965	0.987	benign
Weighted Avg.	0.956	0.039	0.957	0.956	0.956	0.985	

=== Confusion Matrix ===

a	b	<-- classified as
153	5	a = malignant
14	261	b = benign

=== Run information ===

Scheme:weka.classifiers.functions.Logistic -R 1.0E-8 -M -1
Relation: Breast
Instances: 266
Attributes: 10
clump
ucellsize
ucellshape
magadhesion
sepics
bnuclei
bchromatin
normnucl
mitoses
class
Test mode:10-fold cross-validation

=== Classifier model (full training set) ===

Logistic Regression with ridge parameter of 1.0E-8
Coefficients...
Class
Variable malignant
==========

clump 0.7195
ucellsize 0.1651
ucellshape 0.0934
magadhesion 0.7147
sepics 0.501
bnuclei 0.2839
bchromatin 1.9842
normnucl 0.4885
mitoses 3.2358
Intercept -24.0634

Odds Ratios...

Variable	Class malignant
clump	2.0533
ucellsize	1.1795
ucellshape	1.0979
magadhesion	2.0435
sepics	1.6503
bnuclei	1.3283
bchromatin	7.2732
normnucl	1.6299
mitoses	25.4269

Time taken to build model: 0.16 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances	255	95.8647 %
Incorrectly Classified Instances	11	4.1353 %
Kappa statistic	0.9034	
Mean absolute error	0.0409	
Root mean squared error	0.1842	
Relative absolute error	9.5082 %	
Root relative squared error	39.751 %	
Total Number of Instances	266	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.928	0.027	0.939	0.928	0.933	0.981	malignant
	0.973	0.072	0.967	0.973	0.97	0.98	benign
Weighted Avg.	0.959	0.058	0.959	0.959	0.959	0.98	

=== Confusion Matrix ===

a	b	<-- classified as
77	6	a = malignant
5	178	b = benign

=== Re-evaluation on test set ===

User supplied test set
Relation: Breast
Instances: unknown (yet). Reading incrementally
Attributes: 10

=== Summary ===

Correctly Classified Instances	408	94.2263 %
Incorrectly Classified Instances	25	5.7737 %
Kappa statistic	0.8742	
Mean absolute error	0.0599	
Root mean squared error	0.2209	
Total Number of Instances	433	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.899	0.033	0.94	0.899	0.919	0.988	malignant
	0.967	0.101	0.943	0.967	0.955	0.988	benign
Weighted Avg.	0.942	0.076	0.942	0.942	0.942	0.988	

=== Confusion Matrix ===

a	b	<-- classified as
142	16	a = malignant
9	266	b = benign

=== Run information ===

Scheme:weka.classifiers.functions.MultilayerPerceptron -L 0.3 -M 0.2 -N 500 -V 0 -S 0 -E 20 -H a
Relation: Breast
Instances: 266
Attributes: 10
clump
ucellsize
ucellshape
magadhesion
sepics
bnuclei
bchromatin
normnucl
mitoses
class
Test mode:10-fold cross-validation

=== Classifier model (full training set) ===

Sigmoid Node 0
Inputs Weights
Threshold 7.055155276714432
Node 2 -4.337301863948531
Node 3 -3.0161173598673963
Node 4 -0.7982220474723172
Node 5 -6.085042284529997
Node 6 -4.30929039109078
Sigmoid Node 1
Inputs Weights
Threshold -7.051769094486156
Node 2 4.367489741221618
Node 3 3.0135818462756987
Node 4 0.7709079177064214
Node 5 6.084952216490562
Node 6 4.289309962388457
Sigmoid Node 2
Inputs Weights
Threshold -7.060343625080421
Attrib clump -1.3321489951781704
Attrib ucellsize -2.221803038928232
Attrib ucellshape 0.46054977587612655

Attrib magadhesion 0.11109734331744317
Attrib sepics 0.4509101345007964
Attrib bnuclei -0.36464427850075487
Attrib bchromatin -3.8115362484968527
Attrib normnucl -1.113122799678163
Attrib mitoses -4.471968592884483
Sigmoid Node 3
Inputs Weights
Threshold -3.167951207840781
Attrib clump -1.546611979564754
Attrib ucellsize -0.12046672746964626
Attrib ucellshape 1.6813556814481532
Attrib magadhesion -1.3477474996236525
Attrib sepics -0.41418045472018844
Attrib bnuclei 2.9955120549916456
Attrib bchromatin -1.6062368854546754
Attrib normnucl -0.7812300214838922
Attrib mitoses -3.330536526233263
Sigmoid Node 4
Inputs Weights
Threshold -1.832069827023158
Attrib clump -0.5221776395664868
Attrib ucellsize -0.2084443513826472
Attrib ucellshape -0.08818764384123014
Attrib magadhesion 0.06125933748622355
Attrib sepics 0.4618374671709951
Attrib bnuclei 0.27158058176459027
Attrib bchromatin -1.0287805087743764
Attrib normnucl -0.6184457941437396
Attrib mitoses -0.8083815371372266
Sigmoid Node 5
Inputs Weights
Threshold -9.452450717449164
Attrib clump -3.036789601429978
Attrib ucellsize -2.2118472816700887
Attrib ucellshape 3.3404836112210616
Attrib magadhesion -1.7817622651463667
Attrib sepics -1.8821814137866624
Attrib bnuclei -2.627772667188493
Attrib bchromatin -2.589261423029245
Attrib normnucl -0.02366238692326025
Attrib mitoses -5.3786470649415055

Sigmoid Node 6
  Inputs Weights
  Threshold -7.586627165043234
  Attrib clump -1.9088837465055224
  Attrib ucellsize -1.9059819387948327
  Attrib ucellshape 1.86221468566607
  Attrib magadhesion -1.089390921803328
  Attrib sepics -1.1523999000110714
  Attrib bnuclei -1.6643222679285519
  Attrib bchromatin -2.432257581663086
  Attrib normnucl -0.13513124138206933
  Attrib mitoses -4.0148964002917005
Class malignant
  Input
  Node 0
Class benign
  Input
  Node 1

Time taken to build model: 1.52 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances	256	96.2406 %
Incorrectly Classified Instances	10	3.7594 %
Kappa statistic	0.9119	
Mean absolute error	0.0416	
Root mean squared error	0.1828	
Relative absolute error	9.6708 %	
Root relative squared error	39.4424 %	
Total Number of Instances	266	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.928	0.022	0.951	0.928	0.939	0.991	malignant
	0.978	0.072	0.968	0.978	0.973	0.991	benign
Weighted Avg.	0.962	0.057	0.962	0.962	0.962	0.991	

=== Confusion Matrix ===

```
  a   b   <-- classified as
 77   6 |   a = malignant
  4 179 |   b = benign
```

=== Re-evaluation on test set ===

User supplied test set
Relation: Breast
Instances: unknown (yet). Reading incrementally
Attributes: 10

=== Summary ===

Correctly Classified Instances	408	94.2263 %
Incorrectly Classified Instances	25	5.7737 %
Kappa statistic	0.8749	
Mean absolute error	0.0617	
Root mean squared error	0.217	
Total Number of Instances	433	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.911	0.04	0.929	0.911	0.92	0.98	malignant
	0.96	0.089	0.95	0.96	0.955	0.98	benign
Weighted Avg.	0.942	0.071	0.942	0.942	0.942	0.98	

=== Confusion Matrix ===

```
  a   b   <-- classified as
144  14 |   a = malignant
 11 264 |   b = benign
```

=== Run information ===

Scheme:weka.classifiers.trees.J48 -C 0.25 -M 2
Relation: Breast
Instances: 266
Attributes: 10

clump
ucellsize
ucellshape
magadhesion
sepics
bnuclei
bchromatin
normnucl
mitoses
class

Test mode:10-fold cross-validation

=== Classifier model (full training set) ===

J48 pruned tree
---

bchromatin <= 3
| clump <= 6
| | ucellsize <= 2: benign (165.0)
| | ucellsize > 2
| | | ucellsize <= 5: benign (12.0/1.0)
| | | ucellsize > 5: malignant (4.0/1.0)
| clump > 6: malignant (11.0/1.0)
bchromatin > 3
| bnuclei <= 8
| | clump <= 3: benign (2.0)
| | clump > 3
| | | bchromatin <= 4
| | | | mitoses <= 1: benign (4.0/1.0)
| | | | mitoses > 1: malignant (4.0)
| | | bchromatin > 4: malignant (23.0)
| bnuclei > 8: malignant (41.0)

Number of Leaves : 9

Size of the tree : 17

Time taken to build model: 0.06 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	249	93.609 %
Incorrectly Classified Instances	17	6.391 %
Kappa statistic	0.8516	
Mean absolute error	0.0762	
Root mean squared error	0.2489	
Relative absolute error	17.7284 %	
Root relative squared error	53.7172 %	
Total Number of Instances	266	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.904	0.049	0.893	0.904	0.898	0.93	malignant
	0.951	0.096	0.956	0.951	0.953	0.93	benign
Weighted Avg.	0.936	0.082	0.936	0.936	0.936	0.93	

=== Confusion Matrix ===

a	b	<-- classified as
75	8	a = malignant
9	174	b = benign

=== Re-evaluation on test set ===

User supplied test set
Relation: Breast
Instances: unknown (yet). Reading incrementally
Attributes: 10

=== Summary ===

Correctly Classified Instances	406	93.7644 %
Incorrectly Classified Instances	27	6.2356 %
Kappa statistic	0.8634	
Mean absolute error	0.0665	
Root mean squared error	0.2326	
Total Number of Instances	433	

=== Detailed Accuracy By Class ===

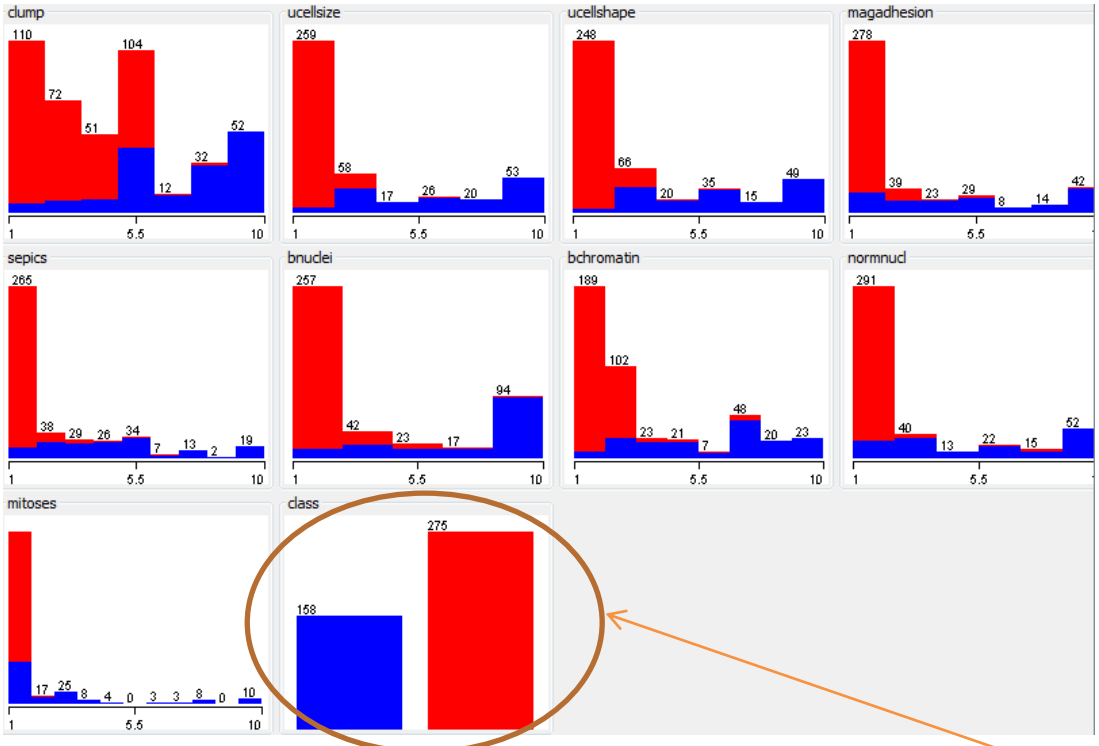
	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.88	0.029	0.946	0.88	0.911	0.962	malignant
	0.971	0.12	0.934	0.971	0.952	0.962	benign
Weighted Avg.	0.938	0.087	0.938	0.938	0.937	0.962	

=== Confusion Matrix ===

a	b	<-- classified as
139	19	a = malignant
8	267	b = benign

Initial Analysis:

The data set being studied has 10 attributes and 433 instances. Given that this dataset is titled Breast, I am assuming it relates to the diagnoses of breast cancer. Based on the class attribute, I am also assuming that 158 cases are not cancer (blue) and 275 cases are cancer (red). Shown below is the graphical distribution of the 10 attributes.



Notice how the attribute Class perfectly classifies the binary response of the diagnoses. The variables initially appear to be continuous.

Observations:

In this exploratory data analysis (EDA) fit and correctly classified instances are major key performance indicators (KPIs). In addition to these KPIs, false positive rate is another major KPI given

that the data is diagnostic in nature. Thus far in this class, we have yet to test our models, which changes with this EDA. The last KPI I will be looking at is the change from the training data to the testing data. The four KPI's listed above will prove to be the delineating factors when assessing the best model.

Naïve Bayes

The term Malignant in medical diagnostics means harmful, where benign is interpreted as non-harmful. Out of the four diagnosis shown below, I care most about B,A which is 1 in the training cross-tab below. This number represents the number of patients that are diagnosed as a tumor being benign but in fact they are falsely benign meaning they are actually malignant. As one can infer, this false diagnostic allows the cancer to grow and decreases the probability of survival for a patient greatly. The rate of False Negative (FN) on the training data is .003% and .01% on the testing data. This is a 233% increase, which is quite high, but overall still amounts to 1% of all patients. Moving forward I would want to test this on a larger data set.

Training

```
  a   b   <-- classified as
 82   1 |   a = malignant
  9 174 |   b = benign
```

Testing

```
  a   b   <-- classified as
 153   5 |   a = malignant
  14 261 |   b = benign
```

The correctly classified instances for training data were 96.2406 and 95.612% for the testing data. Overall this is a .6286% change. In addition, the root mean square error was .1923 for training, and .2043 for testing which is a .06% change. Overall, this is a strong model and I am the most concerned with the false negatives as a KPI moving forward.

Logistic

Logistic Regression is used as a non-linear transformer in the MLP process. The goal of using logistic regression is to linearly analyze data that initially is not linear. This process is done through maximum likelihood estimation.

Training:

```
Correctly Classified Instances         255               95.8647 %
  a   b   <-- classified as
 77   6 |   a = malignant
  5 178 |   b = benign
Root mean squared error                  0.1842
```

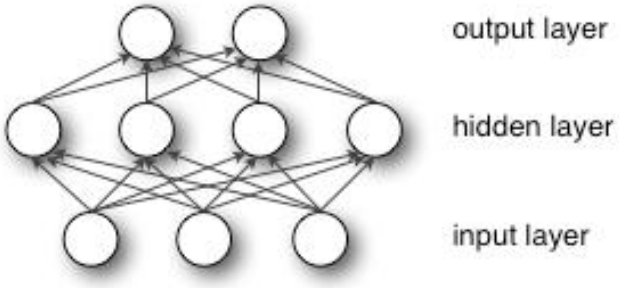
```
=== Re-evaluation on test set ===
Correctly Classified Instances         408               94.2263 %
  a   b   <-- classified as
142  16 |   a = malignant
  9 266 |   b = benign
```

Root mean squared error 0.2209

The rate of False Negative (FN) on the training data is 6/266 = .022 and 16/433= .037 on the testing data. The calculation for change is (.037-.022)/.022 = .681 increase, which is quite high, but overall still amounts to 16/433=.03 or 3% all patients. Comparing this to Naïve Bayes, Naïve Bayes is clearly a better model for this specific KPI. Moving forward I would want to test this on a larger data set. The correctly classified instances for training data were 95.8647 and 94.2263% for the testing data. Overall this is a - 1.6384% change in accuracy. Ideally, one does not want to see accuracy deteriorate. In addition, the root mean square error was .1842 for training, and .2209 for testing which is a (.2209-.1842)/.1842 = 19% increase in errors from the two different models. Overall, I am not impressed with this model compared to the Naïve Bayes. Initially, the root mean square error was lower for Logistic than Naïve Bayes but on the testing data it is larger, which leads me to believe that over fitting might be a small issue.

Multilayer Perceptron

Shown below is an example of a Neural Network that has one hidden layer. Each node had multiple iterations, which is seen through the increase in weight.



=== Classifier model (full training set) ===
Correctly Classified Instances 256 96.2406 %
Root mean squared error 0.1828
a b <-- classified as
77 6 | a = malignant
4 179 | b = benign

=== Re-evaluation on test set ===
Correctly Classified Instances 408 94.2263 %
Root mean squared error 0.217
a b <-- classified as
144 14 | a = malignant
11 264 | b = benign

The rate of False Negative (FN) on the training data is 6/266 = .022 and 14/433= .032 on the testing data. The calculation for change is (.032-.022)/.022 = .454 increase, which is quite high, but overall still amounts to 14/433=.03 or 3% of all patients. Comparing this to Naïve Bayes, Naïve Bayes is clearly a better model for this specific KPI. Moving forward I would want to test this on a larger data set. The correctly classified instances for training data were 96.2406 and 94.2263% for the testing data. Overall this is a - 2.143% change in accuracy. Ideally, one does not want to see accuracy deteriorate. In addition, the root mean square error was .1828 for training, and .217 for testing which is a (.217-.1828)/.1828 = 18% increase in errors from the two different models. Overall, I am not impressed with this model compared to the Naïve Bayes. Initially, the root mean square error was lower for Perceptron than Naïve Bayes but on the testing data it is larger, which leads me to believe that over fitting might be a small issue, if it possible for a neural network. I rank this above logistic based on the fact that the root mean square error is smaller, and the change for false negatives is lower than logistic. This ranking is based on the best desired outcome for the patients being tested.

My current rank is:

- Naïve Bayes
- Multilayer Perceptron
- Logistic

Decision Tree J48

J48 is a top-down approach that separates the example data into subsets (decision tree) and new observations are scored through this tree. The decision tree demonstrataes the top down approach from J48.

```
=== Classifier model (full training set) ===
Number of Leaves  :      9
Size of the tree :       17
Correctly Classified Instances       249          93.609  %
Root mean squared error              0.2489
  a   b   <-- classified as
 75   8 |   a = malignant
  9 174 |   b = benign
```

```
=== Re-evaluation on test set ===
Correctly Classified Instances       406          93.7644 %
Root mean squared error              0.2326
a   b   <-- classified as
 139  19 |   a = malignant
  8 267 |   b = benign
```

The rate of False Negative (FN) on the training data is 8/266 = .03 and 19/433= .043 on the testing data. The calculation for change is (.043-.03)/.03 = .433 increase, which is quite high, but overall

still amounts to 14/433=.03 or 3% of all patients. Comparing this to Naïve Bayes, Naïve Bayes is clearly a better model for this specific KPI. Moving forward I would want to test this on a larger data set.
The correctly classified instances for training data were 93.609 and 93.7644% for the testing data. Overall this is a .1154% increase in accuracy. Ideally, one does not want to see accuracy deteriorate, which speaks highly for this model. In addition, the root mean square error was .2489 for training, and .2326 for testing which is a (.2326-.2489)/.2489 = .0163% decrease in errors from the two different models. Overall, I am very impressed with this model compared to the Naïve Bayes. Initially, the root mean square error was high but dropped slightly, which shows that it fits the data well. I rank this directly below Naïve Bayes based on the fact that the false negatives increased the most and this model has the highest amount of false negatives. This ranking is based on the best desired outcome for the patients being tested.

Conclusion:

Of all four models, I would rank them as shown below:

1. Naïve Bayes
2. Decision Tree
3. Multilayer Perceptron
4. Logistic

My ranking culminated from analyzing 4 metrics that I labeled KPIs. The false negative ratio was my most valuable KPI because this data is based on cancer diagnostics, of which a false negative can kill a person. Across the KPI's Naïve Bayes did the best, with the exception of the accuracy between testing and training data – which Decision tree had the best metric. Multilayer Perceptron and Logistic were the two newer algorithms that have been learned. I was surprised Multilayer Perceptron did not perform better, based on the how intense the logic is behind the algorithm. While I appreciate logistic regression, this was not the best situation or data set for logistic regression. In my opinion, logistic regression really adds value in advertising and marketing datasets. Overall, it was exciting to view this data through the different models with the understanding of the medical consequences.