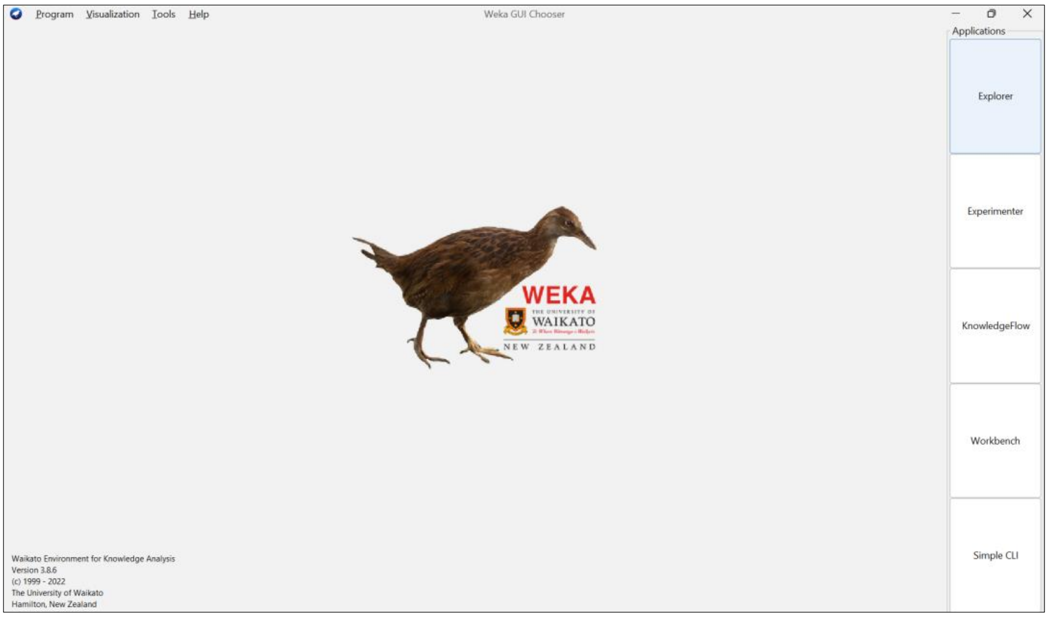
Anirudh Rao
BE21B004

# BT4110 – Computational Biology Lab

## Practical 2

### 1

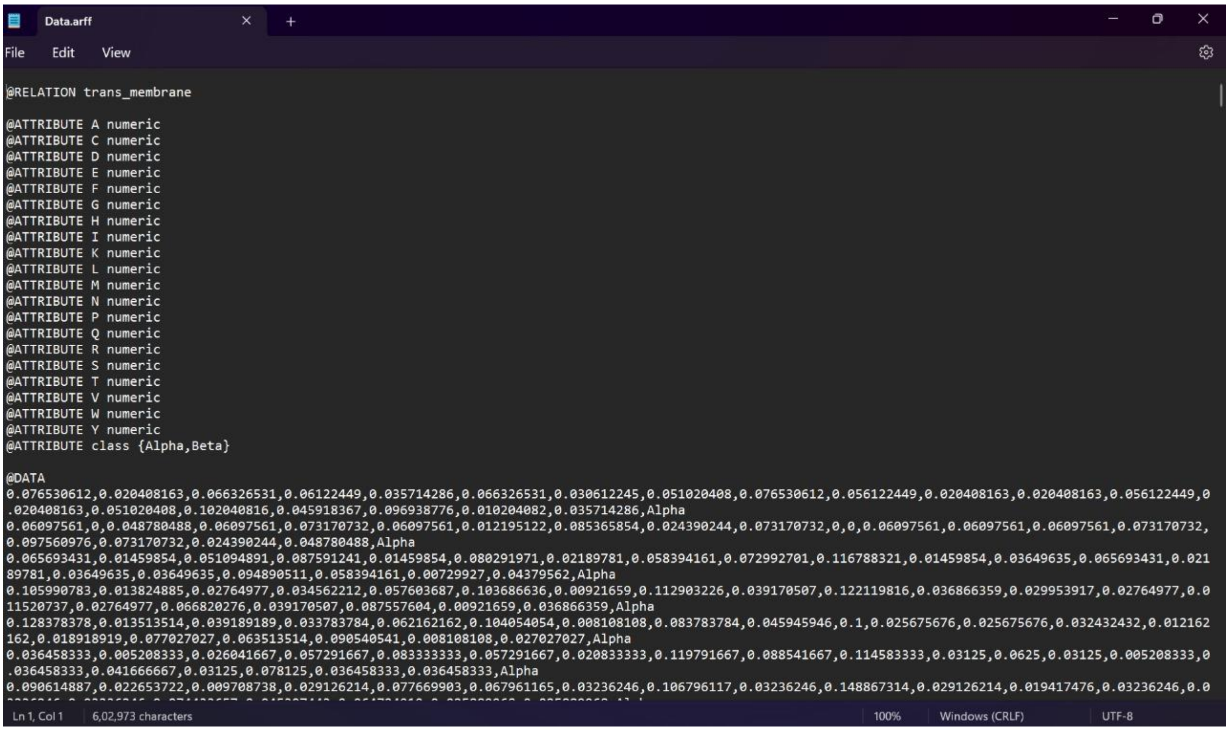
Weka was installed using the link provided.



Waikato Environment for Knowledge Analysis
Version 3.8.6
(c) 1999 - 2022
The University of Waikato
Hamilton, New Zealand

### 2

The input file was prepared in the `.arff` format.

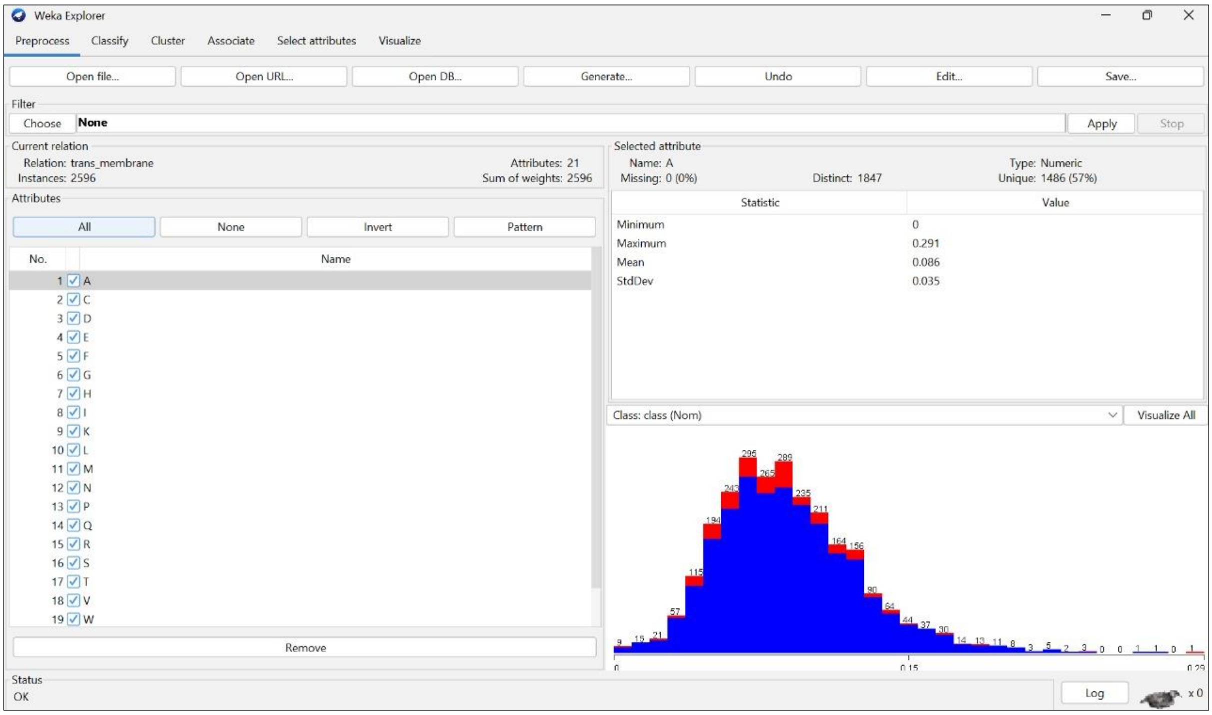


```
@RELATION trans_membrane

@ATTRIBUTE A numeric
@ATTRIBUTE C numeric
@ATTRIBUTE D numeric
@ATTRIBUTE E numeric
@ATTRIBUTE F numeric
@ATTRIBUTE G numeric
@ATTRIBUTE H numeric
@ATTRIBUTE I numeric
@ATTRIBUTE K numeric
@ATTRIBUTE L numeric
@ATTRIBUTE M numeric
@ATTRIBUTE N numeric
@ATTRIBUTE P numeric
@ATTRIBUTE Q numeric
@ATTRIBUTE R numeric
@ATTRIBUTE S numeric
@ATTRIBUTE T numeric
@ATTRIBUTE V numeric
@ATTRIBUTE W numeric
@ATTRIBUTE Y numeric
@ATTRIBUTE class {Alpha,Beta}

@DATA
0.076530612,0.020408163,0.066326531,0.06122449,0.035714286,0.066326531,0.030612245,0.051020408,0.076530612,0.056122449,0.020408163,0.020408163,0.056122449,0
.020408163,0.051020408,0.102040816,0.045918367,0.096938776,0.010204082,0.035714286,Alpha
0.06097561,0,0.048780488,0.06097561,0.073170732,0.06097561,0.012195122,0.085365854,0.024390244,0.073170732,0,0,0.06097561,0.06097561,0.06097561,0.073170732,
0.097560976,0.073170732,0.024390244,0.048780488,Alpha
0.065693431,0.01459854,0.051094891,0.087591241,0.01459854,0.080291971,0.02189781,0.058394161,0.072992701,0.116788321,0.01459854,0.03649635,0.065693431,0.021
89781,0.03649635,0.03649635,0.094890511,0.058394161,0.00729927,0.04379562,Alpha
0.105990783,0.013824885,0.02764977,0.034562212,0.057603687,0.103686636,0.00921659,0.112903226,0.039170507,0.122119816,0.036866359,0.029953917,0.02764977,0.0
11520737,0.02764977,0.066820276,0.039170507,0.087557604,0.00921659,0.036866359,Alpha
0.128378378,0.013513514,0.039189189,0.033783784,0.062162162,0.104054054,0.008108108,0.083783784,0.045945946,0.1,0.025675676,0.025675676,0.032432432,0.012162
162,0.018918919,0.077027027,0.063513514,0.090540541,0.008108108,0.027027027,Alpha
0.036458333,0.005208333,0.026041667,0.057291667,0.083333333,0.057291667,0.020833333,0.119791667,0.088541667,0.114583333,0.03125,0.0625,0.03125,0.005208333,0
.036458333,0.041666667,0.03125,0.078125,0.036458333,0.036458333,Alpha
0.090614887,0.022653722,0.009708738,0.029126214,0.077669903,0.067961165,0.03236246,0.106796117,0.03236246,0.148867314,0.029126214,0.019417476,0.03236246,0.0
```

The data file was opened in Weka.



Filter

Choose None

Current relation

Relation: trans_membrane

Attributes: 21

Instances: 2596

Sum of weights: 2596

Attributes

| No. | | Name |
|---|---|---|
| 1 | ☑ | A |
| 2 | ☑ | C |
| 3 | ☑ | D |
| 4 | ☑ | E |
| 5 | ☑ | F |
| 6 | ☑ | G |
| 7 | ☑ | H |
| 8 | ☑ | I |
| 9 | ☑ | K |
| 10 | ☑ | L |
| 11 | ☑ | M |
| 12 | ☑ | N |
| 13 | ☑ | P |
| 14 | ☑ | Q |
| 15 | ☑ | R |
| 16 | ☑ | S |
| 17 | ☑ | T |
| 18 | ☑ | V |
| 19 | ☑ | W |

Selected attribute

Name: A

Type: Numeric

Missing: 0 (0%)

Distinct: 1847

Unique: 1486 (57%)

| Statistic | Value |
|---|---|
| Minimum | 0 |
| Maximum | 0.291 |
| Mean | 0.086 |
| StdDev | 0.035 |

Class: class (Nom)

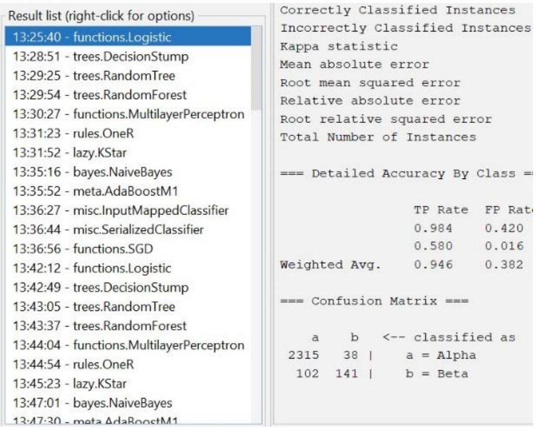
Status

OK

## 4 and 5

10 different machine learning algorithms were used to perform the classification problem and tested for their performance on the training data alone.

| | Training set | | | | | | |
|---|---|---|---|---|---|---|---|
| | **TP** | **TN** | **FP** | **FN** | **Sensitivity** | **Specificity** | **Accuracy** |
| **Logistic** | 2315 | 141 | 102 | 38 | 0.98385 | 0.580247 | 0.94607 |
| **Decision Stump** | 2353 | 0 | 243 | 0 | 1 | 0 | 0.90639 |
| **Random Tree** | 2353 | 242 | 1 | 0 | 1 | 0.995885 | 0.99961 |
| **Random Forest** | 2353 | 242 | 1 | 0 | 1 | 0.995885 | 0.99961 |
| **MultilayerPerceptron** | 2348 | 190 | 53 | 5 | 0.997875 | 0.781893 | 0.97766 |
| **OneR** | 2338 | 42 | 201 | 15 | 0.993625 | 0.17284 | 0.9168 |
| **KStar** | 2353 | 242 | 1 | 0 | 1 | 0.995885 | 0.99961 |
| **Naive Bayes** | 2211 | 185 | 58 | 142 | 0.939652 | 0.761317 | 0.92296 |
| **AdaBoostM1** | 2278 | 131 | 112 | 75 | 0.968126 | 0.539095 | 0.92797 |
| **SGD** | 2310 | 152 | 91 | 43 | 0.981725 | 0.625514 | 0.94838 |



```
Result list (right-click for options)
13:25:40 - functions.Logistic
13:28:51 - trees.DecisionStump
13:29:25 - trees.RandomTree
13:29:54 - trees.RandomForest
13:30:27 - functions.MultilayerPerceptron
13:31:23 - rules.OneR
13:31:52 - lazy.KStar
13:35:16 - bayes.NaiveBayes
13:35:52 - meta.AdaBoostM1
13:36:27 - misc.InputMappedClassifier
13:36:44 - misc.SerializedClassifier
13:36:56 - functions.SGD
13:42:12 - functions.Logistic
13:42:49 - trees.DecisionStump
13:43:05 - trees.RandomTree
13:43:37 - trees.RandomForest
13:44:04 - functions.MultilayerPerceptron
13:44:54 - rules.OneR
13:45:23 - lazy.KStar
13:47:01 - bayes.NaiveBayes
13:47:30 - meta.AdaBoostM1
```

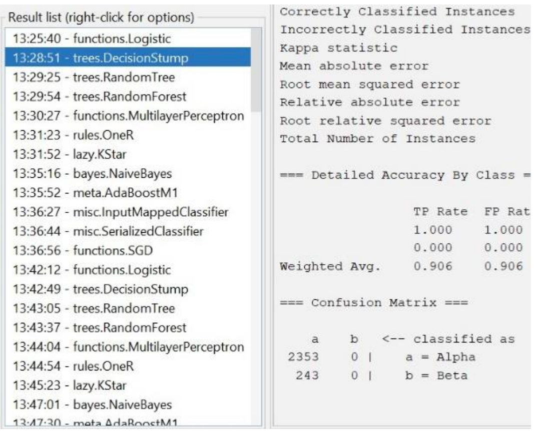
```
Correctly Classified Instances
Incorrectly Classified Instances
Kappa statistic
Mean absolute error
Root mean squared error
Relative absolute error
Root relative squared error
Total Number of Instances

=== Detailed Accuracy By Class =

                 TP Rate  FP Rat
                 0.984    0.420
                 0.580    0.016
Weighted Avg.    0.946    0.382

=== Confusion Matrix ===

    a    b   <-- classified as
 2315   38 |    a = Alpha
  102  141 |    b = Beta
```



```
Result list (right-click for options)
13:25:40 - functions.Logistic
13:28:51 - trees.DecisionStump
13:29:25 - trees.RandomTree
13:29:54 - trees.RandomForest
13:30:27 - functions.MultilayerPerceptron
13:31:23 - rules.OneR
13:31:52 - lazy.KStar
13:35:16 - bayes.NaiveBayes
13:35:52 - meta.AdaBoostM1
13:36:27 - misc.InputMappedClassifier
13:36:44 - misc.SerializedClassifier
13:36:56 - functions.SGD
13:42:12 - functions.Logistic
13:42:49 - trees.DecisionStump
13:43:05 - trees.RandomTree
13:43:37 - trees.RandomForest
13:44:04 - functions.MultilayerPerceptron
13:44:54 - rules.OneR
13:45:23 - lazy.KStar
13:47:01 - bayes.NaiveBayes
13:47:30 - meta.AdaBoostM1
```

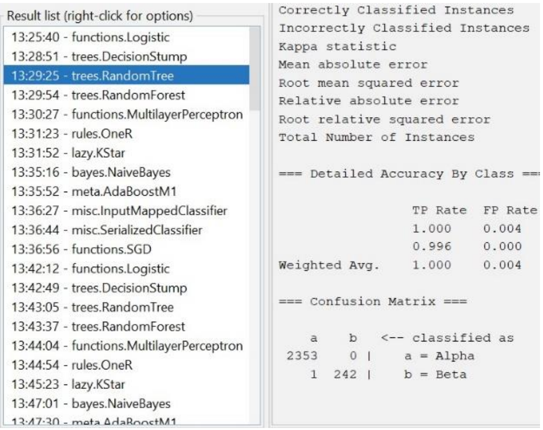
```
Correctly Classified Instances
Incorrectly Classified Instances
Kappa statistic
Mean absolute error
Root mean squared error
Relative absolute error
Root relative squared error
Total Number of Instances

=== Detailed Accuracy By Class =

                 TP Rate  FP Rat
                 1.000    1.000
                 0.000    0.000
Weighted Avg.    0.906    0.906

=== Confusion Matrix ===

    a    b   <-- classified as
 2353    0 |    a = Alpha
  243    0 |    b = Beta
```



```
Result list (right-click for options)
13:25:40 - functions.Logistic
13:28:51 - trees.DecisionStump
13:29:25 - trees.RandomTree
13:29:54 - trees.RandomForest
13:30:27 - functions.MultilayerPerceptron
13:31:23 - rules.OneR
13:31:52 - lazy.KStar
13:35:16 - bayes.NaiveBayes
13:35:52 - meta.AdaBoostM1
13:36:27 - misc.InputMappedClassifier
13:36:44 - misc.SerializedClassifier
13:36:56 - functions.SGD
13:42:12 - functions.Logistic
13:42:49 - trees.DecisionStump
13:43:05 - trees.RandomTree
13:43:37 - trees.RandomForest
13:44:04 - functions.MultilayerPerceptron
13:44:54 - rules.OneR
13:45:23 - lazy.KStar
13:47:01 - bayes.NaiveBayes
13:47:30 - meta.AdaBoostM1
```

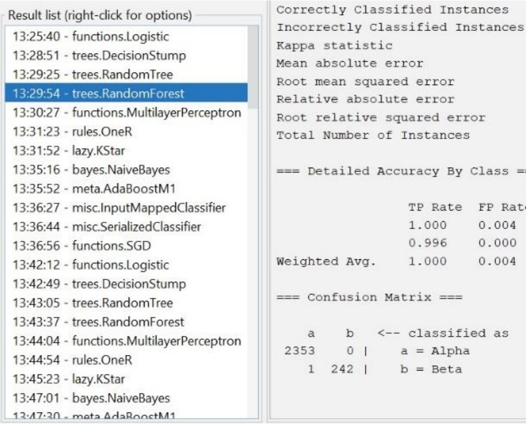
```
Correctly Classified Instances
Incorrectly Classified Instances
Kappa statistic
Mean absolute error
Root mean squared error
Relative absolute error
Root relative squared error
Total Number of Instances

=== Detailed Accuracy By Class ==

                 TP Rate  FP Rate
                 1.000    0.004
                 0.996    0.000
Weighted Avg.    1.000    0.004

=== Confusion Matrix ===

    a    b   <-- classified as
 2353    0 |    a = Alpha
    1  242 |    b = Beta
```



```
Result list (right-click for options)
13:25:40 - functions.Logistic
13:28:51 - trees.DecisionStump
13:29:25 - trees.RandomTree
13:29:54 - trees.RandomForest
13:30:27 - functions.MultilayerPerceptron
13:31:23 - rules.OneR
13:31:52 - lazy.KStar
13:35:16 - bayes.NaiveBayes
13:35:52 - meta.AdaBoostM1
13:36:27 - misc.InputMappedClassifier
13:36:44 - misc.SerializedClassifier
13:36:56 - functions.SGD
13:42:12 - functions.Logistic
13:42:49 - trees.DecisionStump
13:43:05 - trees.RandomTree
13:43:37 - trees.RandomForest
13:44:04 - functions.MultilayerPerceptron
13:44:54 - rules.OneR
13:45:23 - lazy.KStar
13:47:01 - bayes.NaiveBayes
13:47:30 - meta.AdaBoostM1
```

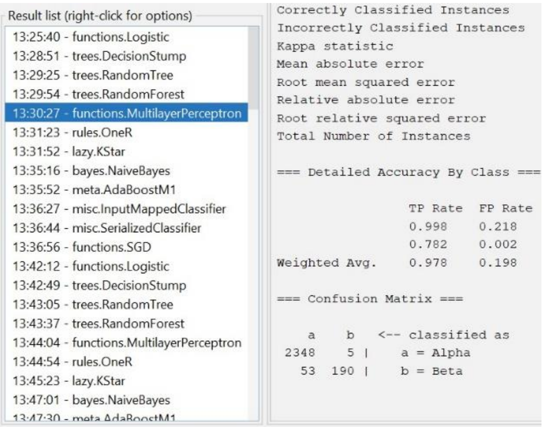
```
Correctly Classified Instances
Incorrectly Classified Instances
Kappa statistic
Mean absolute error
Root mean squared error
Relative absolute error
Root relative squared error
Total Number of Instances

=== Detailed Accuracy By Class =

                 TP Rate  FP Rat
                 1.000    0.004
                 0.996    0.000
Weighted Avg.    1.000    0.004

=== Confusion Matrix ===

    a    b   <-- classified as
 2353    0 |    a = Alpha
    1  242 |    b = Beta
```



```
Result list (right-click for options)
13:25:40 - functions.Logistic
13:28:51 - trees.DecisionStump
13:29:25 - trees.RandomTree
13:29:54 - trees.RandomForest
13:30:27 - functions.MultilayerPerceptron
13:31:23 - rules.OneR
13:31:52 - lazy.KStar
13:35:16 - bayes.NaiveBayes
13:35:52 - meta.AdaBoostM1
13:36:27 - misc.InputMappedClassifier
13:36:44 - misc.SerializedClassifier
13:36:56 - functions.SGD
13:42:12 - functions.Logistic
13:42:49 - trees.DecisionStump
13:43:05 - trees.RandomTree
13:43:37 - trees.RandomForest
13:44:04 - functions.MultilayerPerceptron
13:44:54 - rules.OneR
13:45:23 - lazy.KStar
13:47:01 - bayes.NaiveBayes
13:47:30 - meta.AdaBoostM1
```

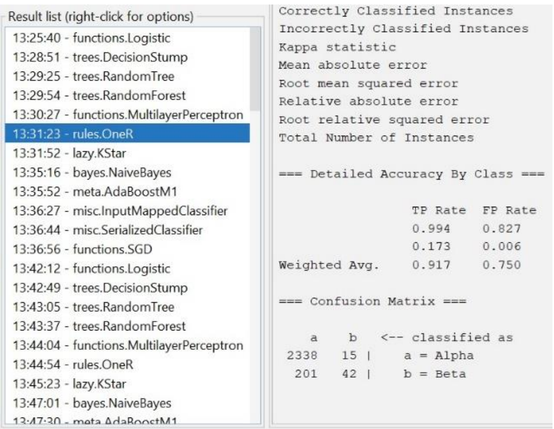
```
Correctly Classified Instances
Incorrectly Classified Instances
Kappa statistic
Mean absolute error
Root mean squared error
Relative absolute error
Root relative squared error
Total Number of Instances

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate
                 0.998    0.218
                 0.782    0.002
Weighted Avg.    0.978    0.198

=== Confusion Matrix ===

    a    b   <-- classified as
 2348    5 |    a = Alpha
   53  190 |    b = Beta
```



```
Result list (right-click for options)
13:25:40 - functions.Logistic
13:28:51 - trees.DecisionStump
13:29:25 - trees.RandomTree
13:29:54 - trees.RandomForest
13:30:27 - functions.MultilayerPerceptron
13:31:23 - rules.OneR
13:31:52 - lazy.KStar
13:35:16 - bayes.NaiveBayes
13:35:52 - meta.AdaBoostM1
13:36:27 - misc.InputMappedClassifier
13:36:44 - misc.SerializedClassifier
13:36:56 - functions.SGD
13:42:12 - functions.Logistic
13:42:49 - trees.DecisionStump
13:43:05 - trees.RandomTree
13:43:37 - trees.RandomForest
13:44:04 - functions.MultilayerPerceptron
13:44:54 - rules.OneR
13:45:23 - lazy.KStar
13:47:01 - bayes.NaiveBayes
13:47:30 - meta.AdaBoostM1
```

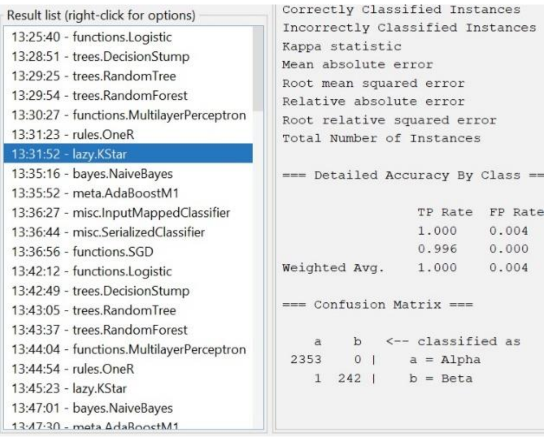
```
Correctly Classified Instances
Incorrectly Classified Instances
Kappa statistic
Mean absolute error
Root mean squared error
Relative absolute error
Root relative squared error
Total Number of Instances

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate
                 0.994    0.827
                 0.173    0.006
Weighted Avg.    0.917    0.750

=== Confusion Matrix ===

    a    b   <-- classified as
 2338   15 |    a = Alpha
  201   42 |    b = Beta
```



```
Result list (right-click for options)
13:25:40 - functions.Logistic
13:28:51 - trees.DecisionStump
13:29:25 - trees.RandomTree
13:29:54 - trees.RandomForest
13:30:27 - functions.MultilayerPerceptron
13:31:23 - rules.OneR
13:31:52 - lazy.KStar
13:35:16 - bayes.NaiveBayes
13:35:52 - meta.AdaBoostM1
13:36:27 - misc.InputMappedClassifier
13:36:44 - misc.SerializedClassifier
13:36:56 - functions.SGD
13:42:12 - functions.Logistic
13:42:49 - trees.DecisionStump
13:43:05 - trees.RandomTree
13:43:37 - trees.RandomForest
13:44:04 - functions.MultilayerPerceptron
13:44:54 - rules.OneR
13:45:23 - lazy.KStar
13:47:01 - bayes.NaiveBayes
13:47:30 - meta.AdaBoostM1
```

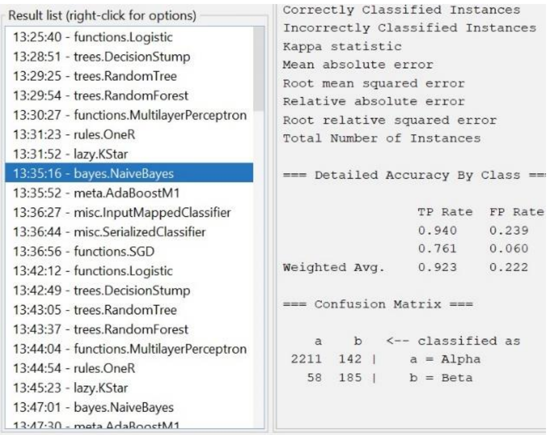
```
Correctly Classified Instances
Incorrectly Classified Instances
Kappa statistic
Mean absolute error
Root mean squared error
Relative absolute error
Root relative squared error
Total Number of Instances

=== Detailed Accuracy By Class ==

                 TP Rate  FP Rate
                 1.000    0.004
                 0.996    0.000
Weighted Avg.    1.000    0.004

=== Confusion Matrix ===

    a    b   <-- classified as
 2353    0 |    a = Alpha
    1  242 |    b = Beta
```



```
Result list (right-click for options)
13:25:40 - functions.Logistic
13:28:51 - trees.DecisionStump
13:29:25 - trees.RandomTree
13:29:54 - trees.RandomForest
13:30:27 - functions.MultilayerPerceptron
13:31:23 - rules.OneR
13:31:52 - lazy.KStar
13:35:16 - bayes.NaiveBayes
13:35:52 - meta.AdaBoostM1
13:36:27 - misc.InputMappedClassifier
13:36:44 - misc.SerializedClassifier
13:36:56 - functions.SGD
13:42:12 - functions.Logistic
13:42:49 - trees.DecisionStump
13:43:05 - trees.RandomTree
13:43:37 - trees.RandomForest
13:44:04 - functions.MultilayerPerceptron
13:44:54 - rules.OneR
13:45:23 - lazy.KStar
13:47:01 - bayes.NaiveBayes
13:47:30 - meta.AdaBoostM1
```

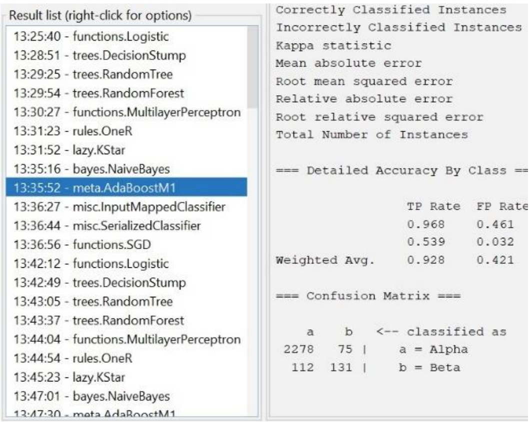
```
Correctly Classified Instances
Incorrectly Classified Instances
Kappa statistic
Mean absolute error
Root mean squared error
Relative absolute error
Root relative squared error
Total Number of Instances

=== Detailed Accuracy By Class ==

                 TP Rate  FP Rate
                 0.940    0.239
                 0.761    0.060
Weighted Avg.    0.923    0.222

=== Confusion Matrix ===

    a    b   <-- classified as
 2211  142 |    a = Alpha
   58  185 |    b = Beta
```



```
13:25:40 - functions.Logistic
13:28:51 - trees.DecisionStump
13:29:25 - trees.RandomTree
13:29:54 - trees.RandomForest
13:30:27 - functions.MultilayerPerceptron
13:31:23 - rules.OneR
13:31:52 - lazy.KStar
13:35:16 - bayes.NaiveBayes
13:35:52 - meta.AdaBoostM1
13:36:27 - misc.InputMappedClassifier
13:36:44 - misc.SerializedClassifier
13:36:56 - functions.SGD
13:42:12 - functions.Logistic
13:42:49 - trees.DecisionStump
13:43:05 - trees.RandomTree
13:43:37 - trees.RandomForest
13:44:04 - functions.MultilayerPerceptron
13:44:54 - rules.OneR
13:45:23 - lazy.KStar
13:47:01 - bayes.NaiveBayes
13:47:30 - meta.AdaBoostM1
```

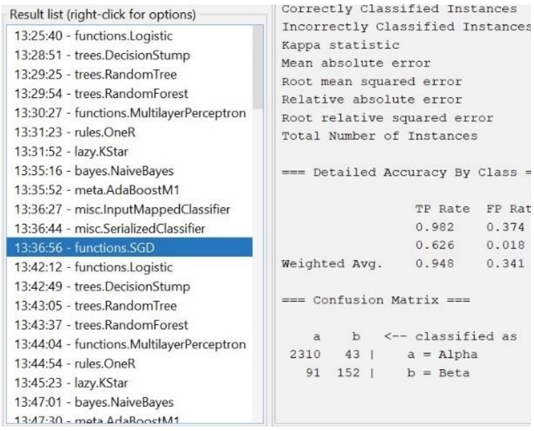
```
Correctly Classified Instances
Incorrectly Classified Instances
Kappa statistic
Mean absolute error
Root mean squared error
Relative absolute error
Root relative squared error
Total Number of Instances

=== Detailed Accuracy By Class ==

                 TP Rate  FP Rate
                 0.968    0.461
                 0.539    0.032
Weighted Avg.    0.928    0.421

=== Confusion Matrix ===

    a    b   <-- classified as
 2278   75 |    a = Alpha
  112  131 |    b = Beta
```



```
13:25:40 - functions.Logistic
13:28:51 - trees.DecisionStump
13:29:25 - trees.RandomTree
13:29:54 - trees.RandomForest
13:30:27 - functions.MultilayerPerceptron
13:31:23 - rules.OneR
13:31:52 - lazy.KStar
13:35:16 - bayes.NaiveBayes
13:35:52 - meta.AdaBoostM1
13:36:27 - misc.InputMappedClassifier
13:36:44 - misc.SerializedClassifier
13:36:56 - functions.SGD
13:42:12 - functions.Logistic
13:42:49 - trees.DecisionStump
13:43:05 - trees.RandomTree
13:43:37 - trees.RandomForest
13:44:04 - functions.MultilayerPerceptron
13:44:54 - rules.OneR
13:45:23 - lazy.KStar
13:47:01 - bayes.NaiveBayes
13:47:30 - meta.AdaBoostM1
```

```
Correctly Classified Instances
Incorrectly Classified Instances
Kappa statistic
Mean absolute error
Root mean squared error
Relative absolute error
Root relative squared error
Total Number of Instances

=== Detailed Accuracy By Class =

                 TP Rate  FP Rat
                 0.982    0.374
                 0.626    0.018
Weighted Avg.    0.948    0.341

=== Confusion Matrix ===

    a    b   <-- classified as
 2310   43 |    a = Alpha
   91  152 |    b = Beta
```

6

| | 5-fold CV | | | | | | |
|---|---|---|---|---|---|---|---|
| | TP | TN | FP | FN | Sensitivity | Specificity | Accuracy |
| **Logistic** | 2308 | 137 | 106 | 45 | 0.980875 | 0.563786 | 0.941834 |
| **Decision Stump** | 2353 | 0 | 243 | 0 | 1 | 0 | 0.906394 |
| **Random Tree** | 2211 | 105 | 138 | 142 | 0.939652 | 0.432099 | 0.892142 |
| **Random Forest** | 2340 | 107 | 136 | 13 | 0.994475 | 0.440329 | 0.942604 |
| **MultilayerPerceptron** | 2278 | 144 | 99 | 75 | 0.968126 | 0.592593 | 0.932974 |
| **OneR** | 2298 | 37 | 206 | 55 | 0.976626 | 0.152263 | 0.899461 |
| **KStar** | 2272 | 127 | 116 | 81 | 0.965576 | 0.522634 | 0.924114 |
| **Naive Bayes** | 2216 | 179 | 64 | 137 | 0.941776 | 0.736626 | 0.922573 |
| **AdaBoostM1** | 2280 | 107 | 136 | 73 | 0.968976 | 0.440329 | 0.919492 |
| **SGD** | 2310 | 143 | 100 | 43 | 0.981725 | 0.588477 | 0.944915 |

| | 10-fold CV | | | | | | |
|---|---|---|---|---|---|---|---|
| | TP | TN | FP | FN | Sensitivity | Specificity | Accuracy |
| **Logistic** | 2307 | 133 | 110 | 46 | 0.98045 | 0.547325 | 0.939908 |
| **Decision Stump** | 2353 | 0 | 243 | 0 | 1 | 0 | 0.906394 |
| **Random Tree** | 2221 | 119 | 124 | 132 | 0.943901 | 0.489712 | 0.901387 |
| **Random Forest** | 2337 | 114 | 129 | 16 | 0.9932 | 0.469136 | 0.944145 |
| **MultilayerPerceptron** | 2291 | 149 | 94 | 62 | 0.973651 | 0.613169 | 0.939908 |
| **OneR** | 2303 | 35 | 208 | 50 | 0.978751 | 0.144033 | 0.900616 |
| **KStar** | 2271 | 127 | 116 | 82 | 0.965151 | 0.522634 | 0.923729 |
| **Naive Bayes** | 2208 | 179 | 64 | 145 | 0.938377 | 0.736626 | 0.919492 |
| **AdaBoostM1** | 2278 | 115 | 128 | 75 | 0.968126 | 0.473251 | 0.921803 |
| **SGD** | 2313 | 138 | 105 | 40 | 0.983 | 0.567901 | 0.944145 |

|  | 20-fold CV |  |  |  |  |  |  |
|---|---|---|---|---|---|---|---|
|  | TP | TN | FP | FN | Sensitivity | Specificity | Accuracy |
| **Logistic** | 2310 | 137 | 106 | 43 | 0.981725 | 0.563786 | 0.942604 |
| **Decision Stump** | 2353 | 0 | 243 | 0 | 1 | 0 | 0.906394 |
| **Random Tree** | 2214 | 115 | 128 | 139 | 0.940926 | 0.473251 | 0.897149 |
| **Random Forest** | 2336 | 110 | 133 | 17 | 0.992775 | 0.452675 | 0.942219 |
| **MultilayerPerceptron** | 2274 | 146 | 97 | 49 | 0.978907 | 0.600823 | 0.943102 |
| **OneR** | 2302 | 23 | 220 | 51 | 0.978326 | 0.09465 | 0.895609 |
| **KStar** | 2272 | 128 | 115 | 81 | 0.965576 | 0.526749 | 0.924499 |
| **Naive Bayes** | 2207 | 181 | 62 | 146 | 0.937952 | 0.744856 | 0.919877 |
| **AdaBoostM1** | 2280 | 120 | 123 | 73 | 0.968976 | 0.493827 | 0.924499 |
| **SGD** | 2313 | 137 | 106 | 40 | 0.983 | 0.563786 | 0.94376 |

|  | 66% split |  |  |  |  |  |  |
|---|---|---|---|---|---|---|---|
|  | TP | TN | FP | FN | Sensitivity | Specificity | Accuracy |
| **Logistic** | 792 | 41 | 36 | 14 | 0.98263 | 0.532468 | 0.943375 |
| **Decision Stump** | 806 | 0 | 77 | 0 | 1 | 0 | 0.912797 |
| **Random Tree** | 763 | 28 | 49 | 43 | 0.94665 | 0.363636 | 0.89581 |
| **Random Forest** | 801 | 32 | 45 | 5 | 0.993797 | 0.415584 | 0.943375 |
| **MultilayerPerceptron** | 791 | 46 | 31 | 15 | 0.98139 | 0.597403 | 0.947905 |
| **OneR** | 788 | 3 | 74 | 18 | 0.977667 | 0.038961 | 0.89581 |
| **KStar** | 780 | 40 | 37 | 26 | 0.967742 | 0.519481 | 0.928652 |
| **Naive Bayes** | 760 | 58 | 19 | 46 | 0.942928 | 0.753247 | 0.926387 |
| **AdaBoostM1** | 784 | 38 | 39 | 22 | 0.972705 | 0.493506 | 0.930917 |
| **SGD** | 797 | 41 | 36 | 9 | 0.988834 | 0.532468 | 0.949037 |

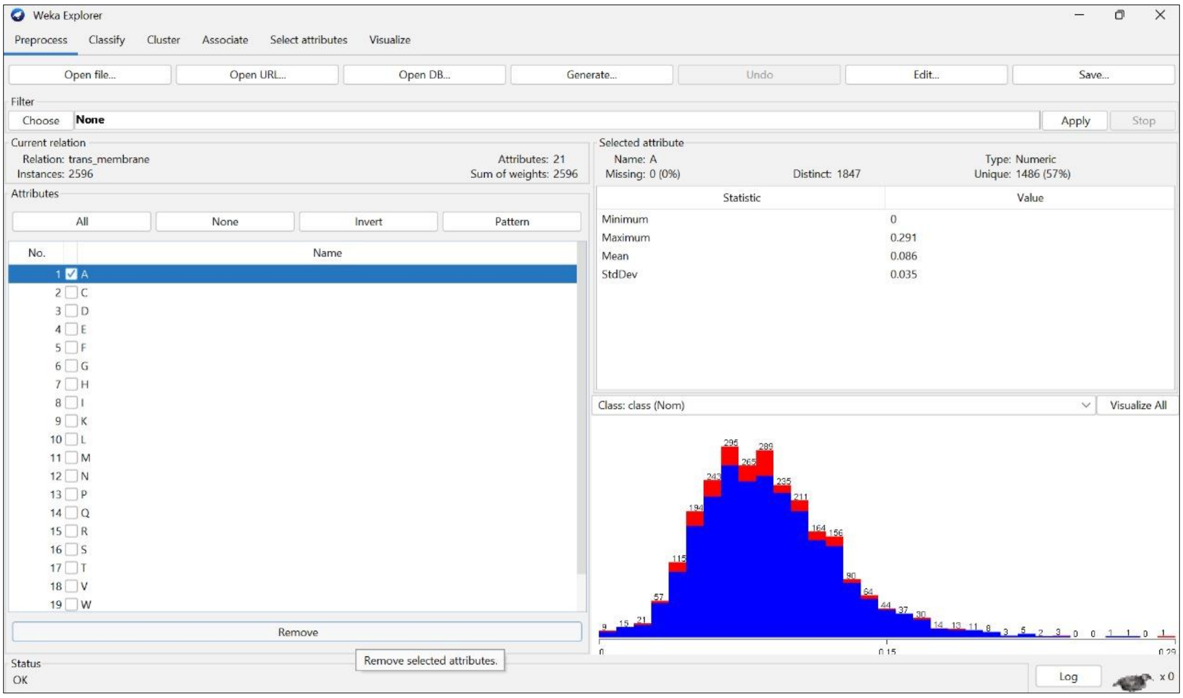
- Most of the models have similar sensitivities and accuracies.
- However, their specificities differ.
- This arises due to the imbalanced nature of the dataset, which has fewer instances of negatives (Beta).
- Thus, using the sensitivity to decide the best model is the appropriate approach.
- Based on this, the **Naive Bayes** classifier has the best performance.

## 7

The best model, Naive Bayes, was tested with different train-test splits.

|  | TP | TN | FP | FN | Sensitivity | Specificity | Accuracy |
|---|---|---|---|---|---|---|---|
| **70%** | 667 | 55 | 18 | 39 | 0.944759 | 0.753425 | 0.926829 |
| **60%** | 883 | 72 | 18 | 65 | 0.931435 | 0.8 | 0.920039 |
| **50%** | 1103 | 94 | 29 | 72 | 0.938723 | 0.764228 | 0.922188 |

Residues were iteratively eliminated to determine their importance. Performance was evaluated using the best model (Naive Bayes) with a 70% train-test split.



Filter

Choose None

Current relation

Relation: trans_membrane

Instances: 2596

Attributes: 21

Sum of weights: 2596

Attributes

| No. | Name |
|---|---|
| 1 | ☑ A |
| 2 | □ C |
| 3 | □ D |
| 4 | □ E |
| 5 | □ F |
| 6 | □ G |
| 7 | □ H |
| 8 | □ I |
| 9 | □ K |
| 10 | □ L |
| 11 | □ M |
| 12 | □ N |
| 13 | □ P |
| 14 | □ Q |
| 15 | □ R |
| 16 | □ S |
| 17 | □ T |
| 18 | □ V |
| 19 | □ W |

Selected attribute

Name: A

Missing: 0 (0%)

Distinct: 1847

Type: Numeric

Unique: 1486 (57%)

| Statistic | Value |
|---|---|
| Minimum | 0 |
| Maximum | 0.291 |
| Mean | 0.086 |
| StdDev | 0.035 |

Class: class (Nom)

Status

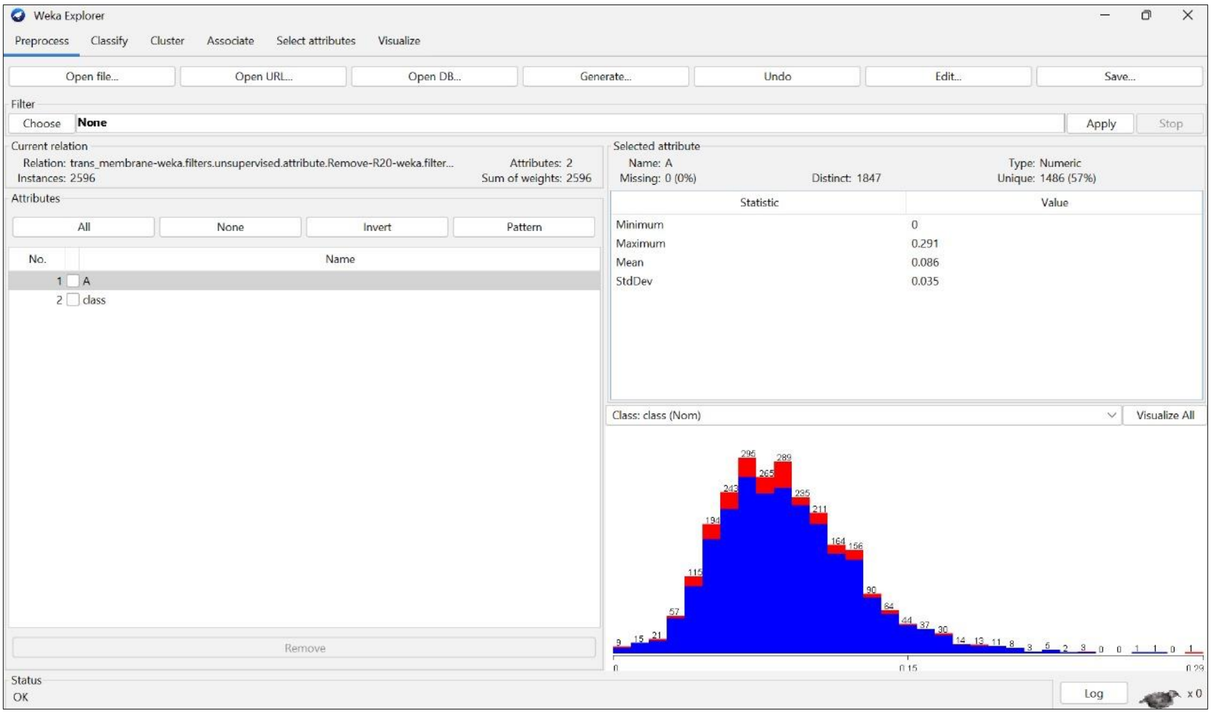
OK

|  | TP | TN | FP | FN | Sensitivity | Specificity | Accuracy |
|---|---|---|---|---|---|---|---|
| **A** | 668 | 55 | 18 | 38 | 0.946176 | 0.753425 | 0.928113 |
| **C** | 669 | 56 | 17 | 37 | 0.947592 | 0.767123 | 0.93068 |
| **D** | 669 | 54 | 19 | 37 | 0.947592 | 0.739726 | 0.928113 |
| **E** | 668 | 55 | 18 | 38 | 0.946176 | 0.753425 | 0.928113 |
| **F** | 675 | 54 | 19 | 31 | 0.956091 | 0.739726 | 0.935815 |
| **G** | 661 | 58 | 15 | 45 | 0.936261 | 0.794521 | 0.922978 |
| **H** | 665 | 55 | 18 | 41 | 0.941926 | 0.753425 | 0.924262 |
| **I** | 669 | 54 | 19 | 37 | 0.947592 | 0.739726 | 0.928113 |
| **K** | 666 | 55 | 18 | 40 | 0.943343 | 0.753425 | 0.925546 |
| **L** | 678 | 54 | 19 | 28 | 0.96034 | 0.739726 | 0.939666 |
| **M** | 672 | 53 | 20 | 34 | 0.951841 | 0.726027 | 0.93068 |
| **N** | 660 | 48 | 25 | 46 | 0.934844 | 0.657534 | 0.908858 |
| **P** | 664 | 54 | 19 | 42 | 0.94051 | 0.739726 | 0.921694 |
| **Q** | 667 | 52 | 21 | 39 | 0.944759 | 0.712329 | 0.922978 |
| **R** | 668 | 54 | 19 | 38 | 0.946176 | 0.739726 | 0.926829 |
| **S** | 663 | 56 | 17 | 43 | 0.939093 | 0.767123 | 0.922978 |
| **T** | 666 | 53 | 20 | 40 | 0.943343 | 0.726027 | 0.922978 |
| **V** | 666 | 56 | 17 | 40 | 0.943343 | 0.767123 | 0.926829 |
| **W** | 668 | 54 | 19 | 38 | 0.946176 | 0.739726 | 0.926829 |
| **Y** | 664 | 59 | 14 | 42 | 0.94051 | 0.808219 | 0.928113 |

- From this, we can see that the sensitivities and accuracies do not change much.
- However, the specificity drops appreciably when Asn (N), Gln (Q), Met (M), or Thr (T) are eliminated.
- N and Q are amino acids with an amide side chain.
- Their compositions may have an important role in discriminating alpha and beta transmembrane proteins.

## 9

The same analysis was performed by using only the composition of a single amino acid as a feature.



Filter
Choose None

Current relation
Relation: trans_membrane-weka.filters.unsupervised.attribute.Remove-R20-weka.filter... Attributes: 2
Instances: 2596 Sum of weights: 2596

Attributes

| No. | | Name |
|---|---|---|
| 1 | □ | A |
| 2 | □ | class |

Selected attribute
Name: A Type: Numeric
Missing: 0 (0%) Distinct: 1847 Unique: 1486 (57%)

| Statistic | Value |
|---|---|
| Minimum | 0 |
| Maximum | 0.291 |
| Mean | 0.086 |
| StdDev | 0.035 |

Class: class (Nom)

Status
OK

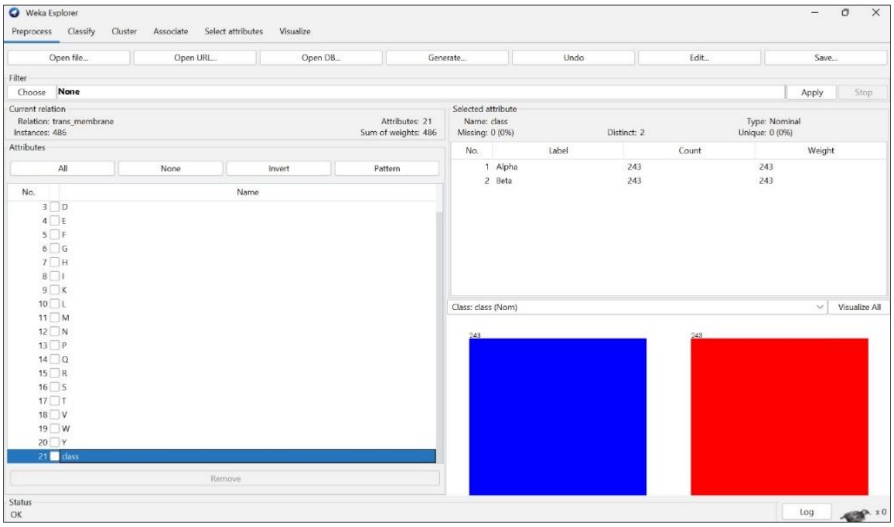
| | TP | TN | FP | FN | Sensitivity | Specificity | Accuracy |
|---|---|---|---|---|---|---|---|
| **A** | 706 | 0 | 73 | 0 | 1 | 0 | 0.90629 |
| **C** | 706 | 0 | 73 | 0 | 1 | 0 | 0.90629 |
| **D** | 700 | 0 | 73 | 6 | 0.991501 | 0 | 0.898588 |
| **E** | 706 | 0 | 73 | 0 | 1 | 0 | 0.90629 |
| **F** | 706 | 0 | 73 | 0 | 1 | 0 | 0.90629 |
| **G** | 696 | 0 | 73 | 10 | 0.985836 | 0 | 0.893453 |
| **H** | 706 | 0 | 73 | 0 | 1 | 0 | 0.90629 |
| **I** | 706 | 0 | 73 | 0 | 1 | 0 | 0.90629 |
| **K** | 706 | 0 | 73 | 0 | 1 | 0 | 0.90629 |
| **L** | 706 | 0 | 73 | 0 | 1 | 0 | 0.90629 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **M** | 706 | 0 | 73 | 0 | 1 | 0 | 0.90629 |
| **N** | 693 | 9 | 64 | 13 | 0.981586 | 0.123288 | 0.901155 |
| **P** | 706 | 0 | 73 | 0 | 1 | 0 | 0.90629 |
| **Q** | 706 | 0 | 73 | 0 | 1 | 0 | 0.90629 |
| **R** | 706 | 0 | 73 | 0 | 1 | 0 | 0.90629 |
| **S** | 706 | 0 | 73 | 0 | 1 | 0 | 0.90629 |
| **T** | 706 | 0 | 73 | 0 | 1 | 0 | 0.90629 |
| **V** | 706 | 0 | 73 | 0 | 1 | 0 | 0.90629 |
| **W** | 706 | 0 | 73 | 0 | 1 | 0 | 0.90629 |
| **Y** | 704 | 0 | 73 | 2 | 0.997167 | 0 | 0.903723 |

From this we can see that the specificity is non-zero only when Asn (N) composition is considered,

### 10 and 11

A balanced dataset was constructed by considering only the first 243 sequences from 'Alpha'.



Filter: None

Current relation
Relation: trans_membrane
Instances: 486
Attributes: 21
Sum of weights: 486

Attributes

| No. | Name |
|---|---|
| 3 | ☐ D |
| 4 | ☐ E |
| 5 | ☐ F |
| 6 | ☐ G |
| 7 | ☐ H |
| 8 | ☐ I |
| 9 | ☐ K |
| 10 | ☐ L |
| 11 | ☐ M |
| 12 | ☐ N |
| 13 | ☐ P |
| 14 | ☐ Q |
| 15 | ☐ R |
| 16 | ☐ S |
| 17 | ☐ T |
| 18 | ☐ V |
| 19 | ☐ W |
| 20 | ☐ Y |
| 21 | ☐ class |

Selected attribute
Name: class
Missing: 0 (0%)
Distinct: 2
Type: Nominal
Unique: 0 (0%)

| No. | Label | Count | Weight |
|---|---|---|---|
| 1 | Alpha | 243 | 243 |
| 2 | Beta | 243 | 243 |

Class: class (Nom)

Status
OK

The performance of the Naive Bayes model was evaluated on this with 5-fold cross-validation.

| | |
|---|---|
| **TP** | 185 |
| **TN** | 209 |
| **FP** | 34 |
| **FN** | 58 |
| **Sensitivity** | 0.761317 |
| **Specificity** | 0.860082 |
| **Accuracy** | 0.8107 |

The model maintains its good performance.