Report: Machine Learning Assignment 2

Extract Input Features

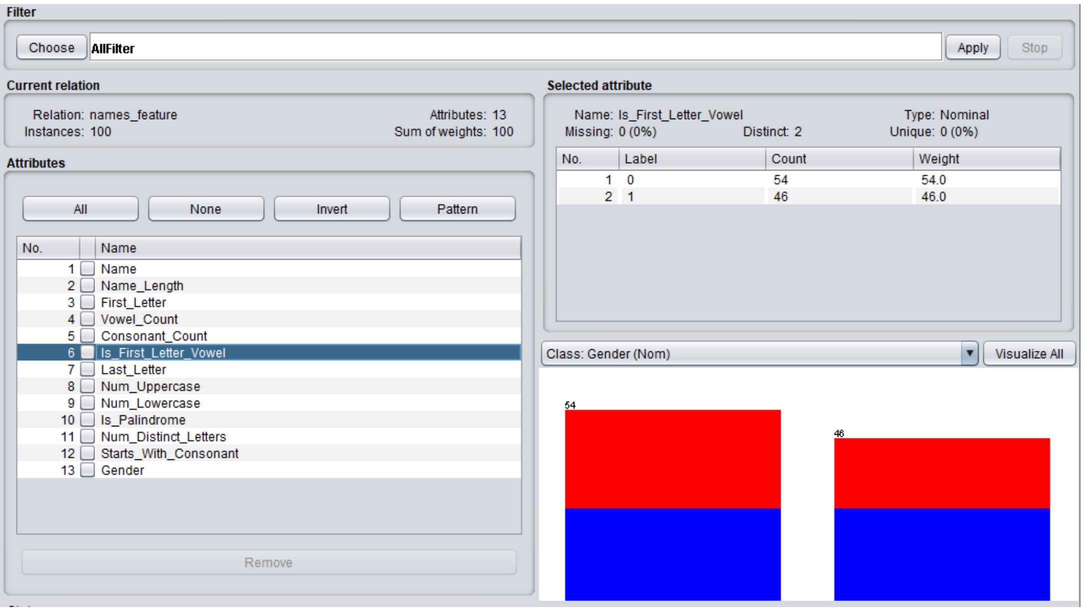
Initially handcrafted features from the "Name" column. Here are a few potential features we could extract:

1. **Name length:** The number of characters in the name.
2. **First letter:** The first character of the name.
3. **Vowel count:** The number of vowels in the name.
4. **Number of consonants**: The number of consonants in the name.
5. **Is first letter a vowel:** Whether the name starts with a vowel (boolean feature).
6. **Name ending:** The last character of the name
7. **Number of uppercase letters:** The count of uppercase letters in the name.
8. **Number of lowercase letters**: The count of lowercase letters in the name.
9. **Is the name a palindrome:** Whether the name reads the same forwards and backwards.
10. **Number of distinct letters:** The count of unique characters in the name.
11. **Name starts with a consonant:** Boolean feature indicating if the name starts with a consonant.
12. **Gender:** Converted numeric output (1 for male, 0 for female).

Interesting Facts

The interesting fact about the data is that, the names start with vowel or consonant belongs to a specific gender which are approximately same as in gender column.

Screenshot:



AllFilter

Relation: names_feature
Instances: 100
Attributes: 13
Sum of weights: 100

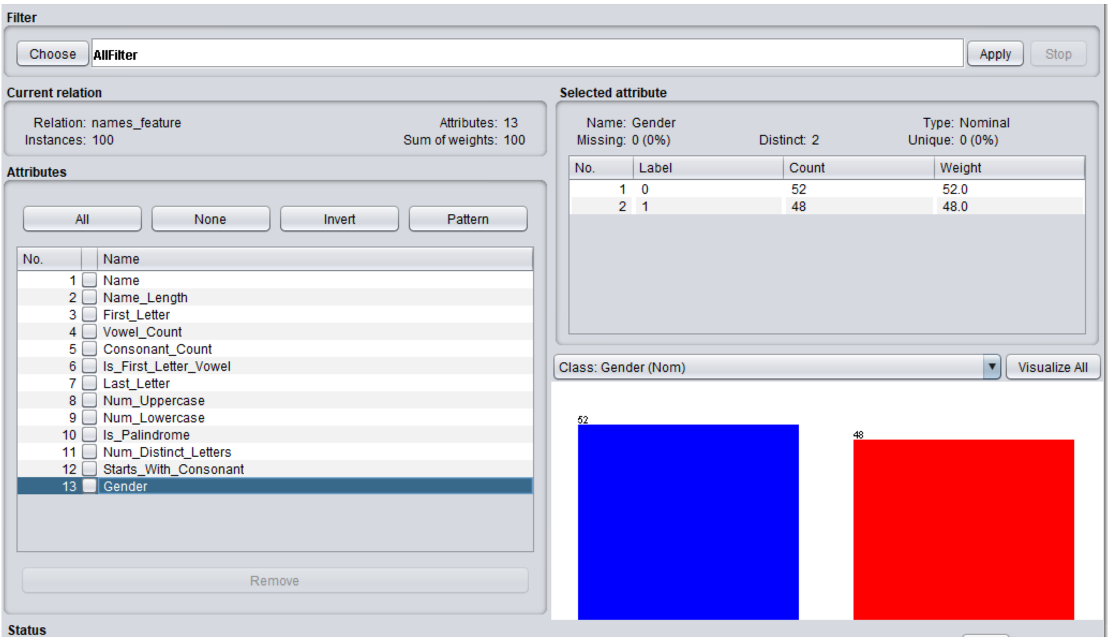
No.		Name
1	□	Name
2	□	Name_Length
3	□	First_Letter
4	□	Vowel_Count
5	□	Consonant_Count
6	□	Is_First_Letter_Vowel
7	□	Last_Letter
8	□	Num_Uppercase
9	□	Num_Lowercase
10	□	Is_Palindrome
11	□	Num_Distinct_Letters
12	□	Starts_With_Consonant
13	□	Gender

Name: Is_First_Letter_Vowel
Type: Nominal
Missing: 0 (0%)
Distinct: 2
Unique: 0 (0%)

No.	Label	Count	Weight
1	0	54	54.0
2	1	46	46.0

Class: Gender (Nom)

Screenshot:



Filter

Choose AllFilter

Current relation

Relation: names_feature
Instances: 100
Attributes: 13
Sum of weights: 100

Attributes

No.	Name
1	Name
2	Name_Length
3	First_Letter
4	Vowel_Count
5	Consonant_Count
6	Is_First_Letter_Vowel
7	Last_Letter
8	Num_Uppercase
9	Num_Lowercase
10	Is_Palindrome
11	Num_Distinct_Letters
12	Starts_With_Consonant
13	Gender

Selected attribute

Name: Gender
Missing: 0 (0%)
Distinct: 2
Type: Nominal
Unique: 0 (0%)

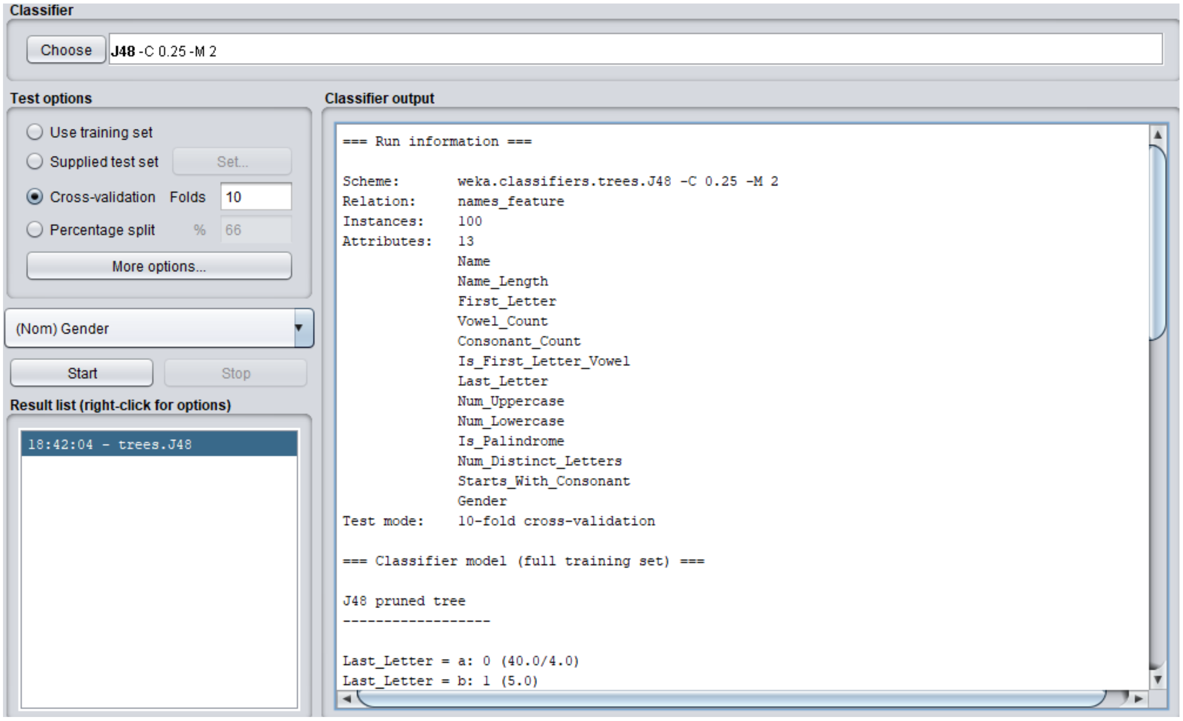
No.	Label	Count	Weight
1	0	52	52.0
2	1	48	48.0

Class: Gender (Nom)

Status

j48 classification algorithm Result

Screenshot 1:



Classifier

Choose J48 -C 0.25 -M 2

Test options

○ Use training set
○ Supplied test set
● Cross-validation Folds 10
○ Percentage split % 66

(Nom) Gender

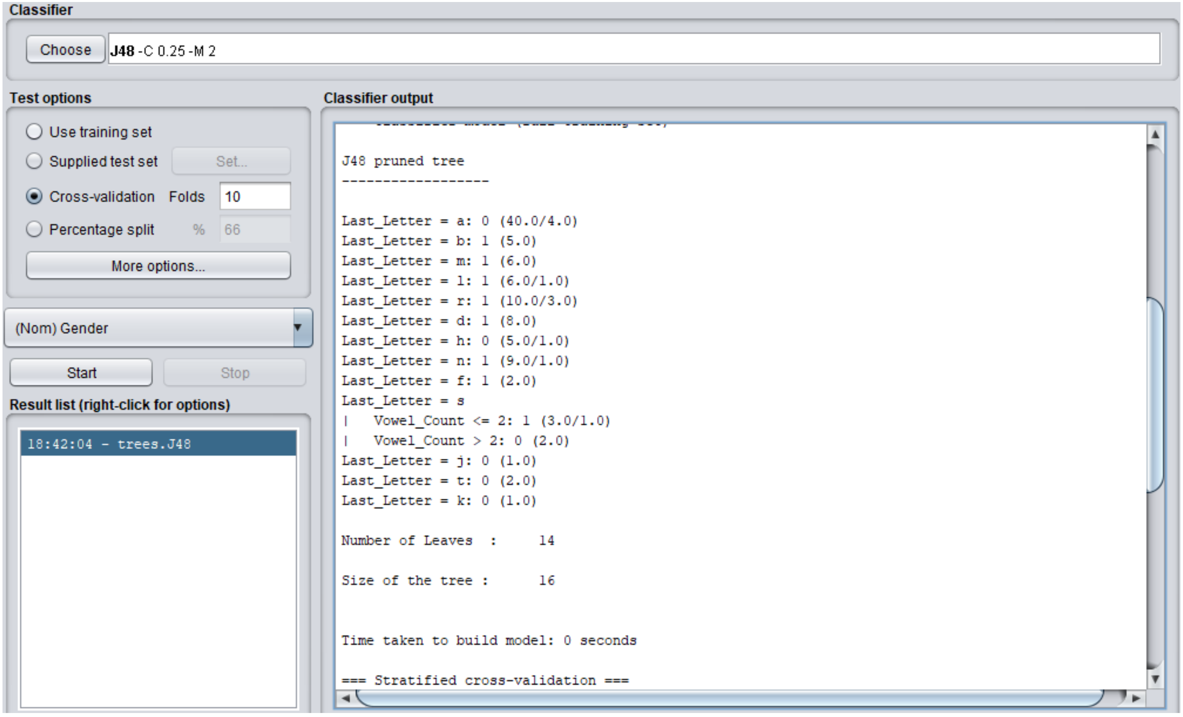
Result list (right-click for options)

18:42:04 - trees.J48

Classifier output

```
=== Run information ===

Scheme:       weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:     names_feature
Instances:    100
Attributes:   13
              Name
              Name_Length
              First_Letter
              Vowel_Count
              Consonant_Count
              Is_First_Letter_Vowel
              Last_Letter
              Num_Uppercase
              Num_Lowercase
              Is_Palindrome
              Num_Distinct_Letters
              Starts_With_Consonant
              Gender
Test mode:    10-fold cross-validation

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

J48 pruned tree
------------------

Last_Letter = a: 0 (40.0/4.0)
Last_Letter = b: 1 (5.0)
```

Screenshot 2:



J48 -C 0.25 -M 2

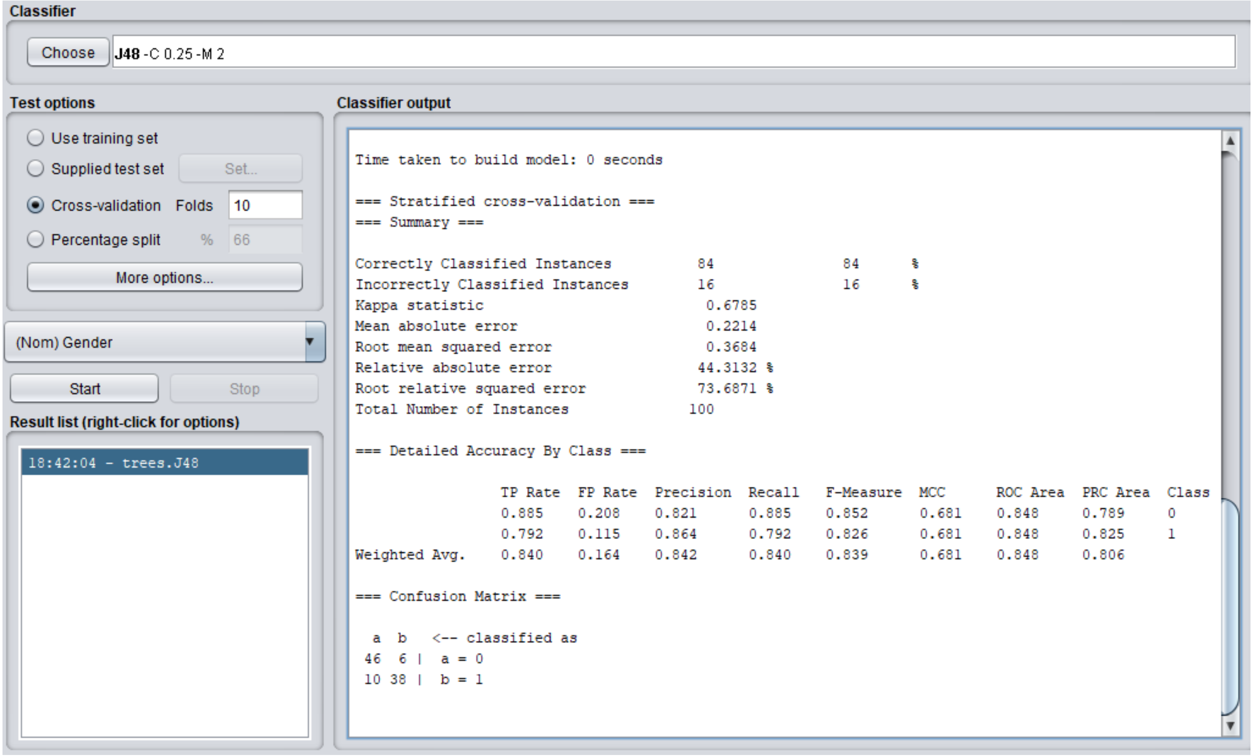
● Cross-validation Folds 10

(Nom) Gender

18:42:04 - trees.J48

```
J48 pruned tree
------------------

Last_Letter = a: 0 (40.0/4.0)
Last_Letter = b: 1 (5.0)
Last_Letter = m: 1 (6.0)
Last_Letter = l: 1 (6.0/1.0)
Last_Letter = r: 1 (10.0/3.0)
Last_Letter = d: 1 (8.0)
Last_Letter = h: 0 (5.0/1.0)
Last_Letter = n: 1 (9.0/1.0)
Last_Letter = f: 1 (2.0)
Last_Letter = s
|   Vowel_Count <= 2: 1 (3.0/1.0)
|   Vowel_Count > 2: 0 (2.0)
Last_Letter = j: 0 (1.0)
Last_Letter = t: 0 (2.0)
Last_Letter = k: 0 (1.0)

Number of Leaves  :     14

Size of the tree :      16


Time taken to build model: 0 seconds

=== Stratified cross-validation ===
```

Screenshot 3:



J48 -C 0.25 -M 2

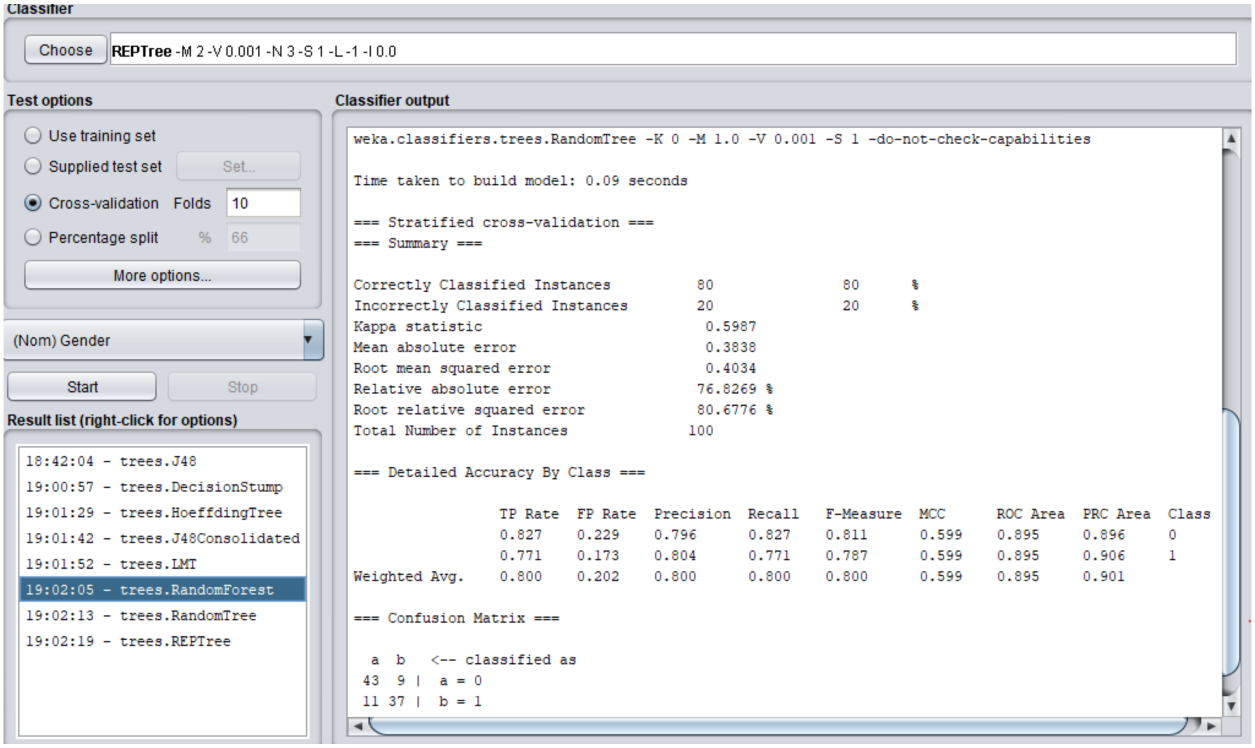
● Cross-validation Folds 10

(Nom) Gender

18:42:04 - trees.J48

```
Time taken to build model: 0 seconds

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

Correctly Classified Instances          84               84      %
Incorrectly Classified Instances        16               16      %
Kappa statistic                          0.6785
Mean absolute error                      0.2214
Root mean squared error                  0.3684
Relative absolute error                 44.3132 %
Root relative squared error             73.6871 %
Total Number of Instances              100

=== Detailed Accuracy By Class ===

                TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                0.885    0.208    0.821      0.885    0.852      0.681    0.848     0.789     0
                0.792    0.115    0.864      0.792    0.826      0.681    0.848     0.825     1
Weighted Avg.   0.840    0.164    0.842      0.840    0.839      0.681    0.848     0.806

=== Confusion Matrix ===

  a  b   <-- classified as
 46  6 |  a = 0
 10 38 |  b = 1
```

Random Forest


Classifier
Choose
Test options
Set...
More options...
Start
Stop
Result list (right-click for options)
18:42:04 - trees.J48
19:00:57 - trees.DecisionStump
19:01:29 - trees.HoeffdingTree
19:01:42 - trees.J48Consolidated
19:01:52 - trees.LMT
19:02:05 - trees.RandomForest
19:02:13 - trees.RandomTree
19:02:19 - trees.REPTree
Classifier output

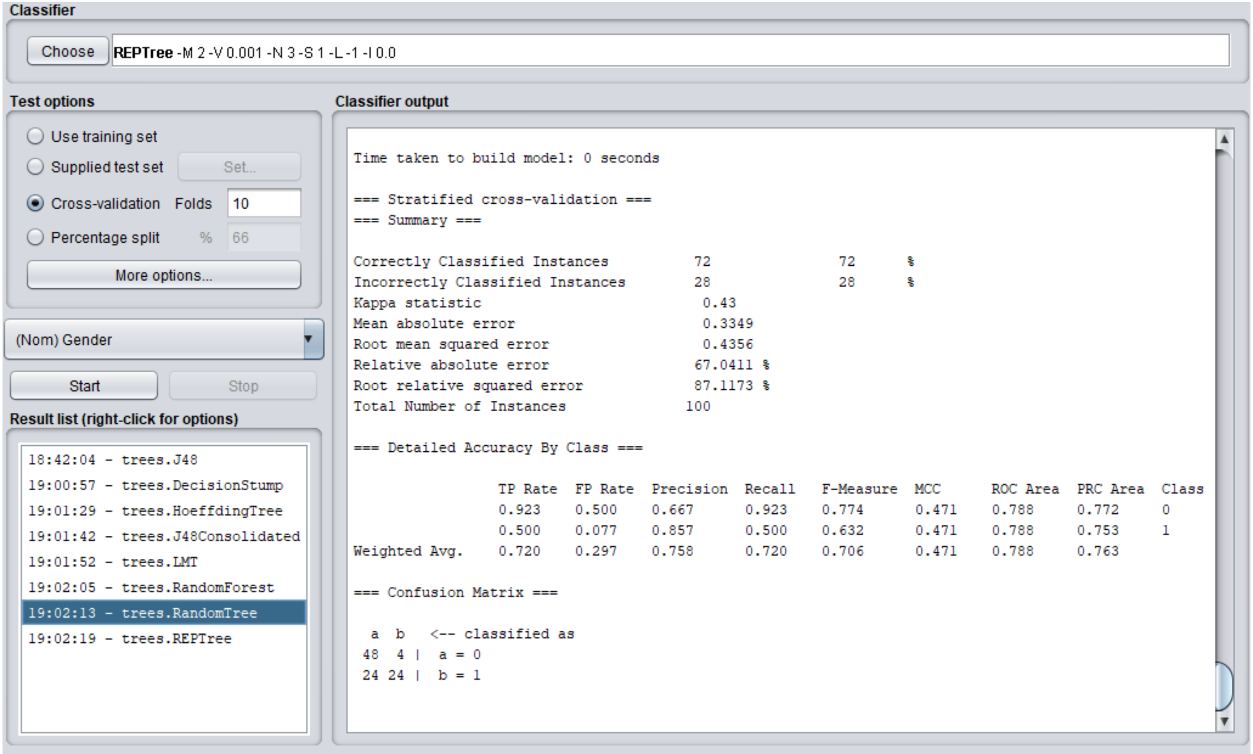
REPTree -M 2 -V 0.001 -N 3 -S 1 -L -1 -I 0.0

○ Use training set
○ Supplied test set
● Cross-validation Folds 10
○ Percentage split % 66

(Nom) Gender

```
weka.classifiers.trees.RandomTree -K 0 -M 1.0 -V 0.001 -S 1 -do-not-check-capabilities

Time taken to build model: 0.09 seconds

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

Correctly Classified Instances          80               80      %
Incorrectly Classified Instances        20               20      %
Kappa statistic                          0.5987
Mean absolute error                      0.3838
Root mean squared error                  0.4034
Relative absolute error                 76.8269 %
Root relative squared error             80.6776 %
Total Number of Instances              100

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.827    0.229    0.796      0.827    0.811      0.599    0.895     0.896     0
                 0.771    0.173    0.804      0.771    0.787      0.599    0.895     0.906     1
Weighted Avg.    0.800    0.202    0.800      0.800    0.800      0.599    0.895     0.901

=== Confusion Matrix ===

  a  b   <-- classified as
 43  9 |  a = 0
 11 37 |  b = 1
```

Random Tree


Classifier
Choose
Test options
Set...
More options...
Start
Stop
Result list (right-click for options)
18:42:04 - trees.J48
19:00:57 - trees.DecisionStump
19:01:29 - trees.HoeffdingTree
19:01:42 - trees.J48Consolidated
19:01:52 - trees.LMT
19:02:05 - trees.RandomForest
19:02:13 - trees.RandomTree
19:02:19 - trees.REPTree
Classifier output

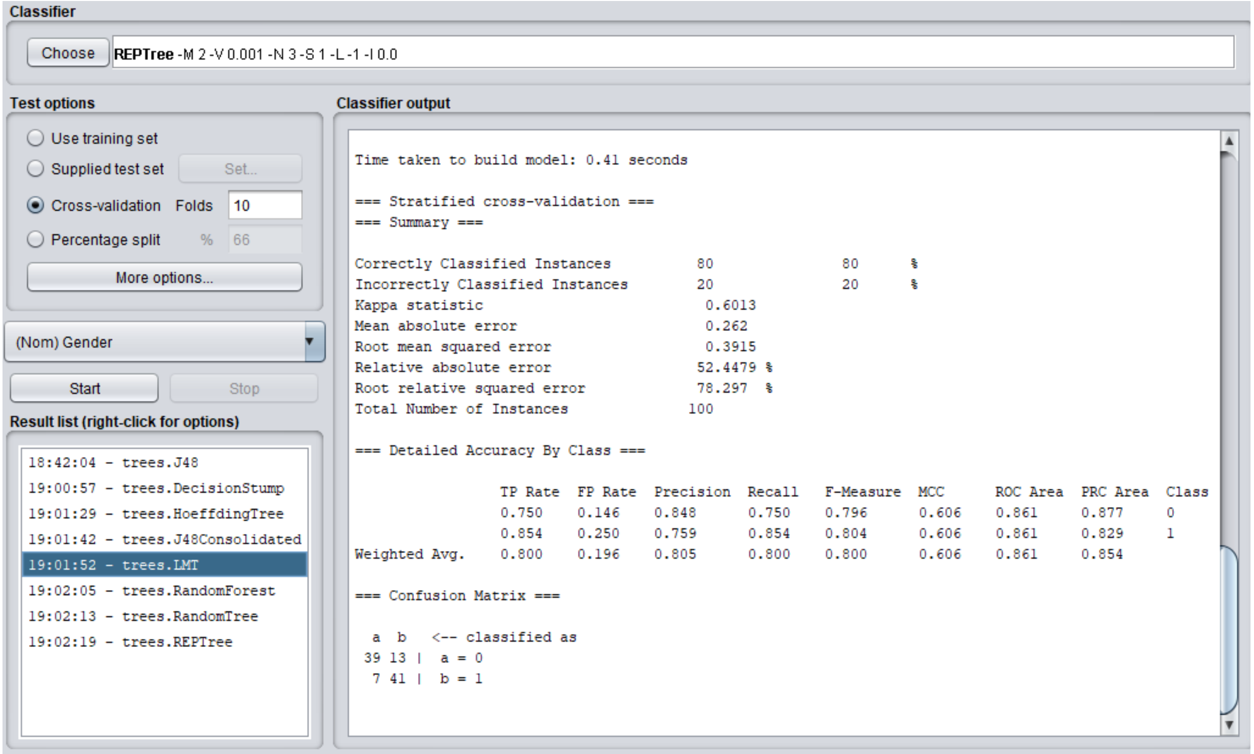
REPTree -M 2 -V 0.001 -N 3 -S 1 -L -1 -I 0.0

○ Use training set
○ Supplied test set
● Cross-validation Folds 10
○ Percentage split % 66

(Nom) Gender

```
Time taken to build model: 0 seconds

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

Correctly Classified Instances          72               72      %
Incorrectly Classified Instances        28               28      %
Kappa statistic                          0.43
Mean absolute error                      0.3349
Root mean squared error                  0.4356
Relative absolute error                 67.0411 %
Root relative squared error             87.1173 %
Total Number of Instances              100

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.923    0.500    0.667      0.923    0.774      0.471    0.788     0.772     0
                 0.500    0.077    0.857      0.500    0.632      0.471    0.788     0.753     1
Weighted Avg.    0.720    0.297    0.758      0.720    0.706      0.471    0.788     0.763

=== Confusion Matrix ===

  a  b   <-- classified as
 48  4 |  a = 0
 24 24 |  b = 1
```

LMT



REPTree -M 2 -V 0.001 -N 3 -S 1 -L -1 -I 0.0

○ Use training set
○ Supplied test set
● Cross-validation Folds 10
○ Percentage split % 66

(Nom) Gender

Result list (right-click for options)

```
18:42:04 - trees.J48
19:00:57 - trees.DecisionStump
19:01:29 - trees.HoeffdingTree
19:01:42 - trees.J48Consolidated
19:01:52 - trees.LMT
19:02:05 - trees.RandomForest
19:02:13 - trees.RandomTree
19:02:19 - trees.REPTree
```

```
Time taken to build model: 0.41 seconds

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

Correctly Classified Instances          80               80      %
Incorrectly Classified Instances        20               20      %
Kappa statistic                          0.6013
Mean absolute error                      0.262
Root mean squared error                  0.3915
Relative absolute error                 52.4479 %
Root relative squared error             78.297  %
Total Number of Instances              100

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.750    0.146    0.848      0.750    0.796      0.606    0.861     0.877     0
                 0.854    0.250    0.759      0.854    0.804      0.606    0.861     0.829     1
Weighted Avg.    0.800    0.196    0.805      0.800    0.800      0.606    0.861     0.854

=== Confusion Matrix ===

  a  b   <-- classified as
 39 13 |  a = 0
  7 41 |  b = 1
```

Experience

Working with the standard machine learning pipeline was an insightful process that highlighted the importance of each step in building a reliable model. The journey began with data preprocessing, where I manually extracted relevant features from the dataset, transforming raw text (names) into meaningful numerical representations. This step was crucial, as the quality and relevance of input features directly influenced the classifier's performance. One most important thing that I learned during this process is conversion of CSV file into ARFF using Weka and manually updating the ARFF file to make it compatible with j48 classification algorithm. Once the data was prepared, I loaded it into WEKA, a user-friendly tool for machine learning experiments. WEKA's interface made it easy to visualize the data, explore attribute relationships, and run various classification algorithms like J48. The process of fine-tuning the model, analyzing the results, and observing how different features affected the predictions was a learning experience in balancing accuracy and model complexity. Though achieving 100% accuracy wasn't possible, the exercise demonstrated how iterative improvements, through feature engineering and model adjustment, can lead to better performance in real-world scenarios.