Machine Learning Assignment
Analysis of Name Characteristics Using J48 and other Classifiers

Submitted By:

Name:	Saira Shairi
Registratin No:	SP24-RCS-023

Submitted To:

Dr. Muhammad Sharjeel

Date:06-10-2024

Introduction:

In this assignment, I explore the application of machine learning classifiers to a dataset of names, and try to understand understanding how various name characteristics influence classification outcomes. The dataset consists of names labeled with +,-,y,n etc outcomes, and main objective is to identify patterns in the names that could be used for classification. For this, I manually extracted several features from the names, like second letter is a vowel, whether the name starts or ends with a vowel, and whether the length of the name is even or odd. For this analysis I have used classifiers mainly the J48 decision tree and some other classifier. This assignment offering an opportunity to better understand the full machine learning pipeline, from feature extraction to model evaluation.

Features and Dataset Explanation:

To effectively classify names based on their characteristics, I manually extracted several features from the dataset.The features extracted include:

1. 2nd alphabet is vowel
2. Length is even or odd
3. Name of girl or boy
4. Name Starts with a vowel

After extracting these features, I converted the dataset into ARFF format to ensure compatibility with the WEKA machine learning environment.

Methodology:

The dataset was formatted into ARFF, making it compatible with WEKA, a popular machine learning tool. I applied different classifiers: the J48 decision tree and the Logistic Model Tree (LMT).

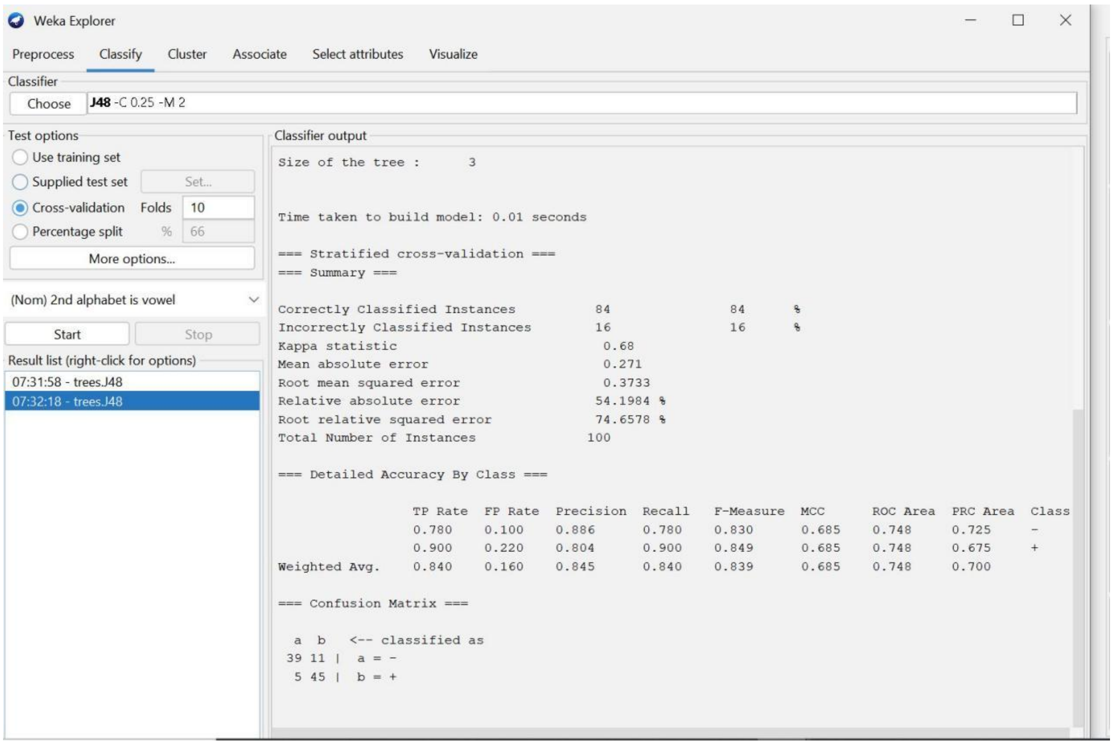
Classifier Outputs for Individual Attributes

I have included screenshots of the outputs obtained after running the classifiers on each attribute individually.

1. 2nd Alphabet is vowel:

Correctly classifed instances 84 %

Incorrectly classified instances 16%



J48 -C 0.25 -M 2

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

(Nom) 2nd alphabet is vowel

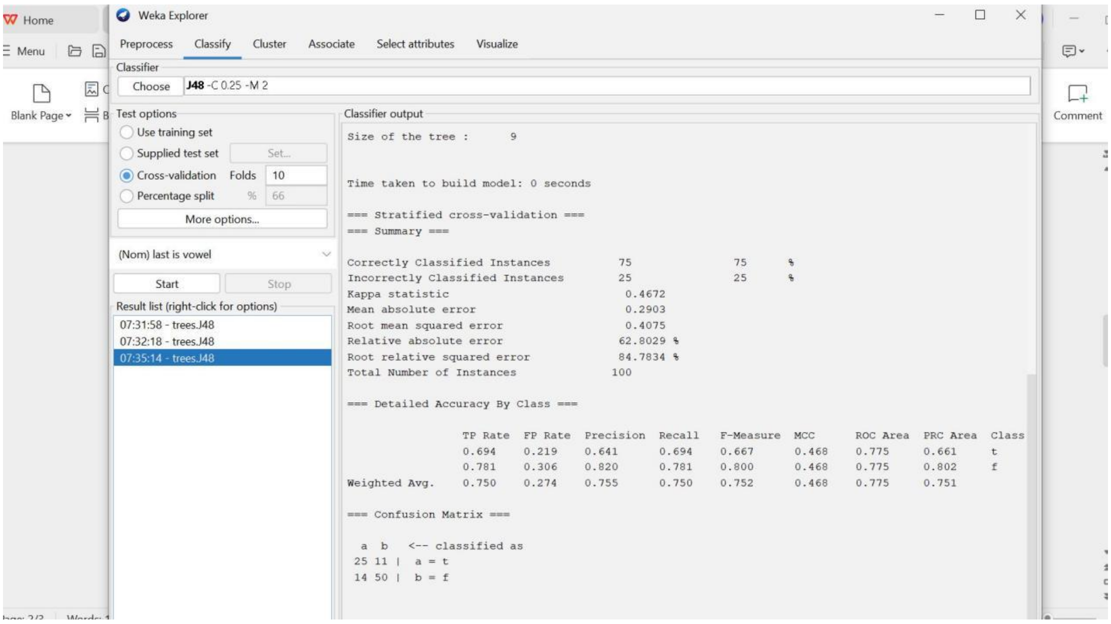
Result list (right-click for options)
07:31:58 - trees.J48
07:32:18 - trees.J48

```
Size of the tree :      3


Time taken to build model: 0.01 seconds

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

Correctly Classified Instances          84               84      %
Incorrectly Classified Instances        16               16      %
Kappa statistic                          0.68
Mean absolute error                      0.271
Root mean squared error                  0.3733
Relative absolute error                 54.1984 %
Root relative squared error             74.6578 %
Total Number of Instances              100

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.780    0.100    0.886      0.780    0.830      0.685    0.748     0.725     -
                 0.900    0.220    0.804      0.900    0.849      0.685    0.748     0.675     +
Weighted Avg.    0.840    0.160    0.845      0.840    0.839      0.685    0.748     0.700

=== Confusion Matrix ===

  a  b   <-- classified as
 39 11 |  a = -
  5 45 |  b = +
```

2. Last is vowel

Correctly classifed instances 75 %

Incorrectly classified instances 25%



Classifier

Choose J48 -C 0.25 -M 2

Test options

○ Use training set

○ Supplied test set

● Cross-validation Folds 10

○ Percentage split % 66

(Nom) last is vowel

Result list (right-click for options)

07:31:58 - trees.J48

07:32:18 - trees.J48

07:35:14 - trees.J48

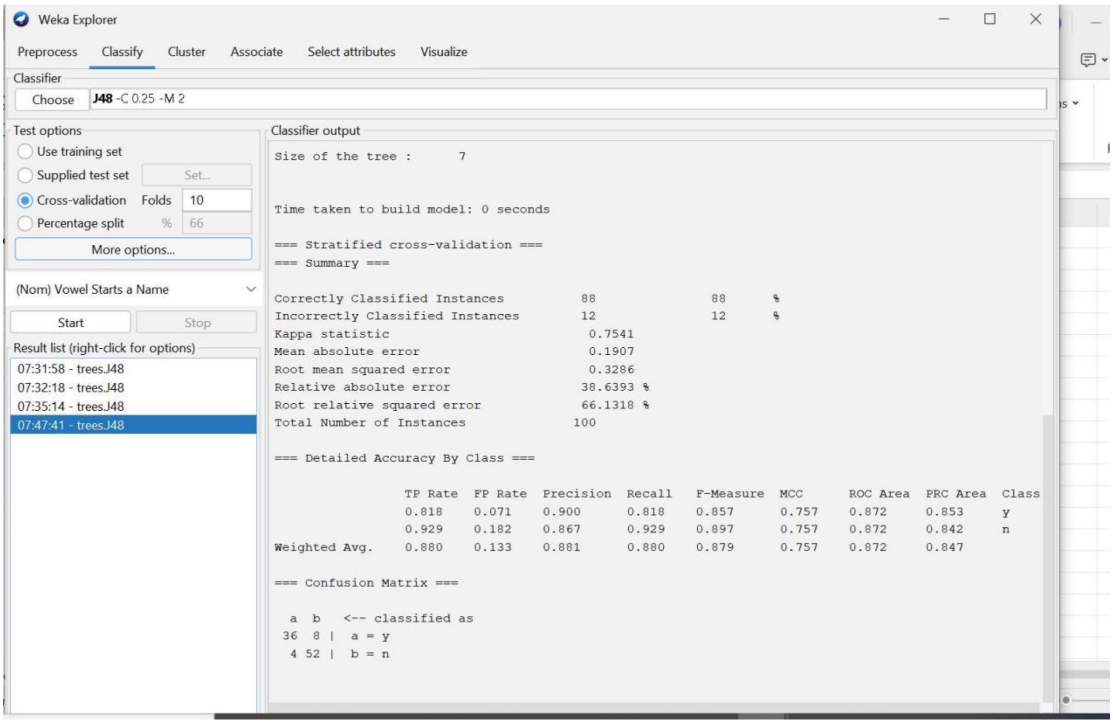
Classifier output

```
Size of the tree :      9


Time taken to build model: 0 seconds

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

Correctly Classified Instances          75               75      %
Incorrectly Classified Instances        25               25      %
Kappa statistic                          0.4672
Mean absolute error                      0.2903
Root mean squared error                  0.4075
Relative absolute error                 62.8029 %
Root relative squared error             84.7834 %
Total Number of Instances              100

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.694    0.219    0.641      0.694    0.667      0.468    0.775     0.661     t
                 0.781    0.306    0.820      0.781    0.800      0.468    0.775     0.802     f
Weighted Avg.    0.750    0.274    0.755      0.750    0.752      0.468    0.775     0.751

=== Confusion Matrix ===

  a  b   <-- classified as
 25 11 |  a = t
 14 50 |  b = f
```

3. Vowel starts a name

Correctly classifed instances 88 %

Incorrectly classified instances 12%



Classifier

Choose J48 -C 0.25 -M 2

Test options

○ Use training set

○ Supplied test set

● Cross-validation Folds 10

○ Percentage split % 66

(Nom) Vowel Starts a Name

Result list (right-click for options)

07:31:58 - trees.J48

07:32:18 - trees.J48

07:35:14 - trees.J48

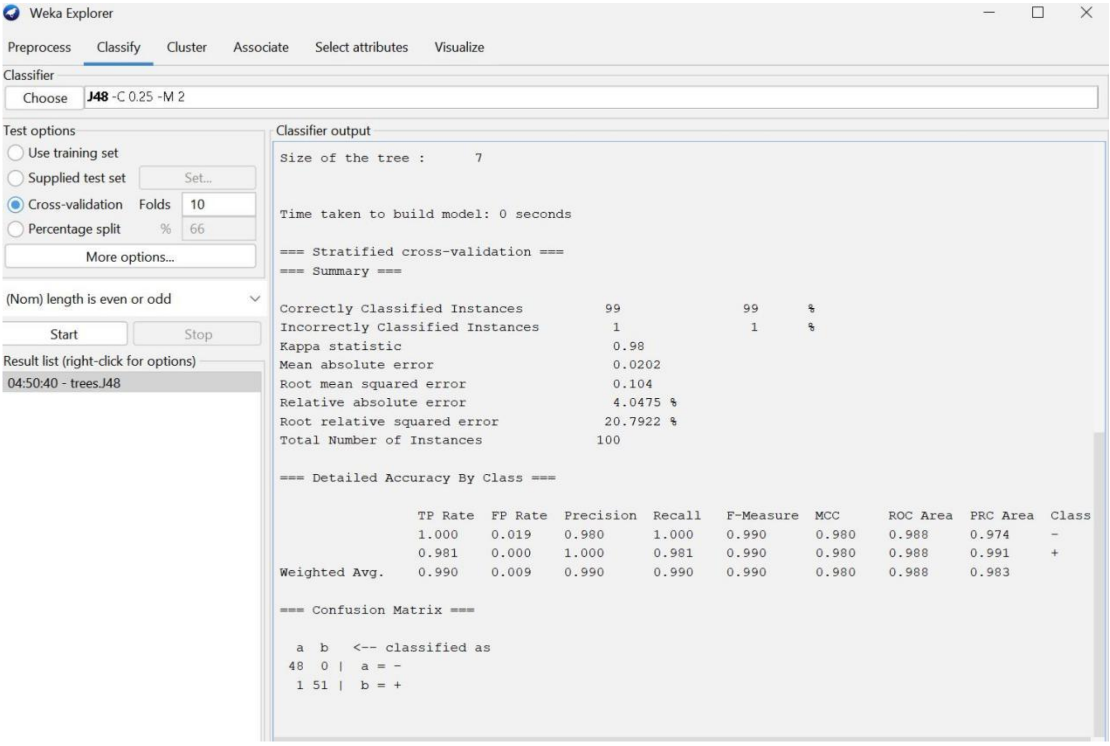
07:47:41 - trees.J48

Classifier output

```
Size of the tree :      7


Time taken to build model: 0 seconds

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

Correctly Classified Instances          88               88      %
Incorrectly Classified Instances        12               12      %
Kappa statistic                          0.7541
Mean absolute error                      0.1907
Root mean squared error                  0.3286
Relative absolute error                 38.6393 %
Root relative squared error             66.1318 %
Total Number of Instances              100

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.818    0.071    0.900      0.818    0.857      0.757    0.872     0.853     y
                 0.929    0.182    0.867      0.929    0.897      0.757    0.872     0.842     n
Weighted Avg.    0.880    0.133    0.881      0.880    0.879      0.757    0.872     0.847

=== Confusion Matrix ===

  a  b   <-- classified as
 36  8 |  a = y
  4 52 |  b = n
```

4. length is even or odd

Correctly classifed instances 99 %
Incorrectly classified instances 1%



J48 -C 0.25 -M 2

● Cross-validation Folds 10

(Nom) length is even or odd

04:50:40 - trees.J48

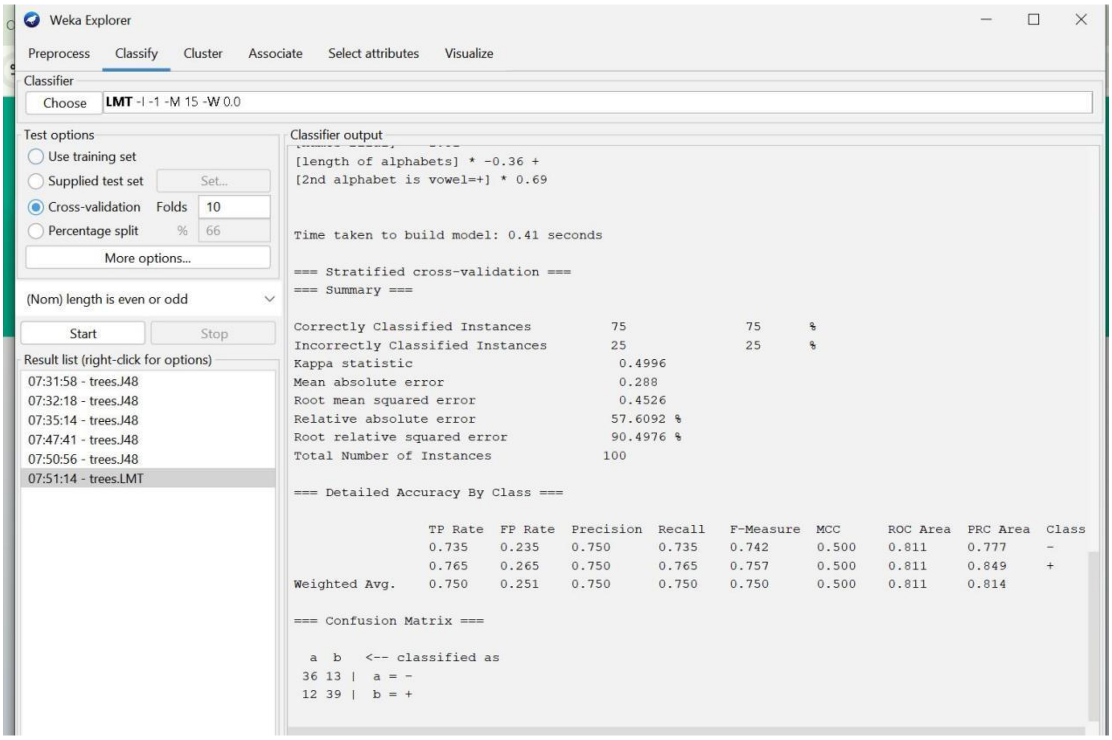
```
Size of the tree :      7


Time taken to build model: 0 seconds

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

Correctly Classified Instances          99               99      %
Incorrectly Classified Instances         1                1      %
Kappa statistic                          0.98
Mean absolute error                      0.0202
Root mean squared error                  0.104
Relative absolute error                  4.0475 %
Root relative squared error             20.7922 %
Total Number of Instances              100

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000    0.019    0.980      1.000    0.990      0.980    0.988     0.974     -
                 0.981    0.000    1.000      0.981    0.990      0.980    0.988     0.991     +
Weighted Avg.    0.990    0.009    0.990      0.990    0.990      0.980    0.988     0.983

=== Confusion Matrix ===

  a  b   <-- classified as
 48  0 |  a = -
  1 51 |  b = +
```

Results:

In this task I have applied J48 decision tree algorithm on my data-set. From my dataset when I choose attribute **(length is even or odd)** it shows high accuracy 99% and show 1 incorrectly classified instance. When I try to correct that 1 incorrectly classified instance by rechecking my dataset and find out that I have make mistake at one place then I correct that mistake and again apply algorithm the **interesting thing** which I have noticed is that its accuracy drop to 70's range from 99%. So, I observe this shift in accuracy shows the potential overfitting in the previous model when it was just dependent on a single attribute, as well as how sensitive decision trees may be to slight changes in data.

Logistic Model Tree (LMT) classifier Result:

I have applied other algorithms also but here I share (LMT) results



Choose **LMT** -I -1 -M 15 -W 0.0

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

(Nom) length is even or odd

Result list (right-click for options)
07:31:58 - trees.J48
07:32:18 - trees.J48
07:35:14 - trees.J48
07:47:41 - trees.J48
07:50:56 - trees.J48
07:51:14 - trees.LMT

```
[length of alphabets] * -0.36 +
[2nd alphabet is vowel=+] * 0.69


Time taken to build model: 0.41 seconds

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

Correctly Classified Instances          75               75      %
Incorrectly Classified Instances        25               25      %
Kappa statistic                          0.4996
Mean absolute error                      0.288
Root mean squared error                  0.4526
Relative absolute error                 57.6092 %
Root relative squared error             90.4976 %
Total Number of Instances              100

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.735    0.235    0.750      0.735    0.742      0.500    0.811     0.777     -
                 0.765    0.265    0.750      0.765    0.757      0.500    0.811     0.849     +
Weighted Avg.    0.750    0.251    0.750      0.750    0.750      0.500    0.811     0.814

=== Confusion Matrix ===

  a  b   <-- classified as
 36 13 |  a = -
 12 39 |  b = +
```

1.length is even or odd

Correctly classifed instances 75 %
Incorrectly classified instances 25%

Result:

I have also applied Logistic Model Tree (LMT) classifier on same above dataset attribute **(length is even or odd)** .The model achieved 75% accuracy, correctly classifying 75 instances and misclassifying 25 .Intrestingly J48 initially achieved 99% accuracy, but this was reduced to the 70s range after fixing a mislabeled instance, revealing signs of overfitting. In contrast, LMT yielded a consistent 75% accuracy, demonstrating its stability when applied to the same set of features. This process highlighted the sensitivity of decision trees to data variations and the importance of selecting robust classifiers for reliable performance.

Conclusion:

In this assignment I have examined the application of machine learning classifiers to analyze name characteristics through feature extraction. The classifiers used demonstrated varying levels of accuracy, emphasizing the complexities of model training and the influence of data quality on performance. The experience underscored the significance of careful feature selection and data preprocessing in the machine learning pipeline.