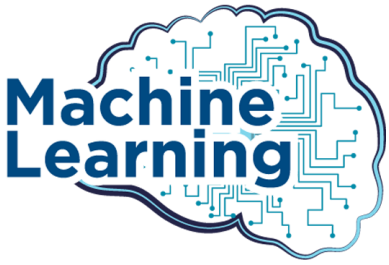
COMSATS University
Islamabad
Lahore Campus

Machine Learning Assignment #2

Analysis of Name Characteristics Using J48 and Other Classifiers



Submitted By:
Name: Hadia Anwer
Registration No: FA24-RCS-003

Submitted To:
Dr. Muhammad Sharjeel

To facilitate effective name classification, I manually extracted a set of relevant features from the dataset. These features include:

1. 2nd alphabet is vowel

2. Length is even or odd

3. Length of name

4. Numbers of vowels

5. Vowel starts a name

Following feature extraction, I transformed the dataset into ARFF (Attribute-Relation File Format) to ensure seamless compatibility with WEKA's machine learning environment, enabling efficient classification model development.

Classifier Outputs for Individual Attributes

A. When Length is even or odd

This indicates that the classifier performed quite well, correctly identifying most of the relevant results, But did not identify 100% which was expected of it.

```
Number of Leaves  :     3

Size of the tree :      5


Time taken to build model: 0.01 seconds

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

Correctly Classified Instances          80               80      %
Incorrectly Classified Instances        20               20      %
Kappa statistic                          0.2372
Mean absolute error                      0.303
Root mean squared error                  0.4081
Relative absolute error                 90.2649 %
Root relative squared error            100.112  %
Total Number of Instances              100

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               0.949     0.762     0.824      0.949    0.882      0.267    0.616     0.828     No
               0.238     0.051     0.556      0.238    0.333      0.267    0.616     0.343     Yes
Weighted Avg.  0.800     0.613     0.768      0.800    0.767      0.267    0.616     0.726

=== Confusion Matrix ===

  a  b   <-- classified as
 75  4 |  a = No
 16  5 |  b = Yes
```

B. Number of Vowels

This suggests that the classifier is doing reasonably well but could benefit from improvements in reducing false positives.

```
Test mode:    split 66.0% train, remainder test

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

M5 pruned model tree:
(using smoothed linear models)
LM1 (100/12.309%)

LM num: 1
Number of Vowels =
        1 * Name=Inaya,Munira,Zeenat,Hadiyya,Khalida,Raniya,Firdous,Aqeel,Asiyah,Jamila,Amira,Khadijah,Na
        + 0.6667 * Name=Muneeb,Afeefah,Maleeha,Aatika,Naeema,Aakifa,Aneesa,AbdulAlim,Saadiya,Maymouna,Ade
        + 0.3333 * Name=Aatika,Naeema,Aakifa,Aneesa,AbdulAlim,Saadiya,Maymouna,Adeela
        + 2

Number of Rules : 1

Time taken to build model: 0.01 seconds

=== Evaluation on test split ===

Time taken to test model on test split: 0 seconds

=== Summary ===

Correlation coefficient                  0
Mean absolute error                      0.6471
Root mean squared error                  0.8044
Relative absolute error                107.0796 %
Root relative squared error            121.6772 %
Total Number of Instances               34
```

C. Starts with a vowel

With this precision, the classifier is making as many correct predictions as incorrect ones, indicating an average performance.

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

ZeroR predicts class value: n

Time taken to build model: 0 seconds

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

Correctly Classified Instances          56               56      %
Incorrectly Classified Instances        44               44      %
Kappa statistic                          0
Mean absolute error                      0.4935
Root mean squared error                  0.4969
Relative absolute error                100      %
Root relative squared error            100      %
Total Number of Instances              100

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.000    0.000    ?          0.000    ?          ?        0.451     0.418     y
                 1.000    1.000    0.560      1.000    0.718      ?        0.451     0.539     n
Weighted Avg.    0.560    0.560    ?          0.560    ?          ?        0.451     0.486

=== Confusion Matrix ===

  a  b   <-- classified as
  0 44 |  a = y
  0 56 |  b = n
```

D. 2nd alphabet is vowel

This feature would be labeled as the **best possible outcome(Magical Feature)**, as it suggests the classifier only identifies truly relevant instances without any mistakes in this context.

```
Number of Leaves  :     2

Size of the tree :      3


Time taken to build model: 0 seconds

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

Correctly Classified Instances         100               100      %
Incorrectly Classified Instances         0                 0      %
Kappa statistic                          1
Mean absolute error                      0
Root mean squared error                  0
Relative absolute error                  0      %
Root relative squared error              0      %
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    0.000    1.000      1.000    1.000      1.000    1.000     1.000     No
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     Yes
Weighted Avg.    1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000

=== Confusion Matrix ===

  a  b   <-- classified as
 50  0 |  a = No
  0 50 |  b = Yes
```

E. When feature is length of the Name

This is a very high precision, meaning the classifier is highly accurate and making very few false positive errors.

```
              OutPut
Test mode:    10-fold cross-validation

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

M5 pruned model tree:
(using smoothed linear models)
LM1 (100/0%)

LM num: 1
Length =
        1 * Name=Abida,Anees,Adeeb,Abeer,Habib,Areej,Uqbah,Inaya,Munir,Aqeel,Imran,Bilal,Hamza,Iffat,Ajma
        + 1 * Name=Aneesa,Naeema,Luqman,Bashir,Raniya,Adeela,Ismail,Fareed,Shakir,Arshad,Asiyah,Bushra,Sh
        + 1 * Name=Hussein,Afeefah,Khalida,Khawlah,Saadiya,Maleeha,Mujtaba,Hadiyya,Sharifa,Ghaliya,Firdou
        + 1 * Name=Khadijah,Maymouna,AbdulAlim,BadrUddin
        + 1 * Name=AbdulAlim,BadrUddin
        + 4

Number of Rules : 1

Time taken to build model: 0.02 seconds

=== Cross-validation ===
=== Summary ===

Correlation coefficient                  0.0593
Mean absolute error                      0.94
Root mean squared error                  1.2247
Relative absolute error                102.9949 %
Root relative squared error            109.603  %
Total Number of Instances              100
```

Write a paragraph about your experience of working with the standard ML pipeline in your own words.

Working with the standard machine learning pipeline in WEKA provided a seamless and intuitive experience. After loading the ARFF file into the workbench, I was able to explore the dataset's characteristics, including attribute distributions and summary statistics. Running the J48 classification algorithm yielded insightful results, showcasing the model's performance and identifying key predictors. WEKA's user-friendly interface facilitated easy navigation and configuration of the algorithm's parameters. The visualizations and output helped identify areas of improvement, such as handling missing values and feature optimization. Overall, WEKA's streamlined pipeline enabled efficient experimentation, allowing me to focus on interpreting results and refining the model.