COMSATS University
Islamabad
Lahore Campus

**Course Code: CSC668**

# Assignment #2

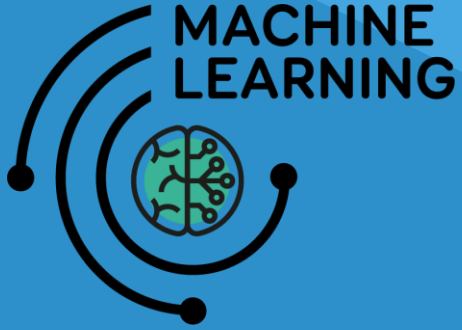
Machine Learning Experiments in WEKA

Oct 06, 2024

Azka Khalid | FA24-RCS-002

Submitted to

Dr. Muhammad Sharjeel

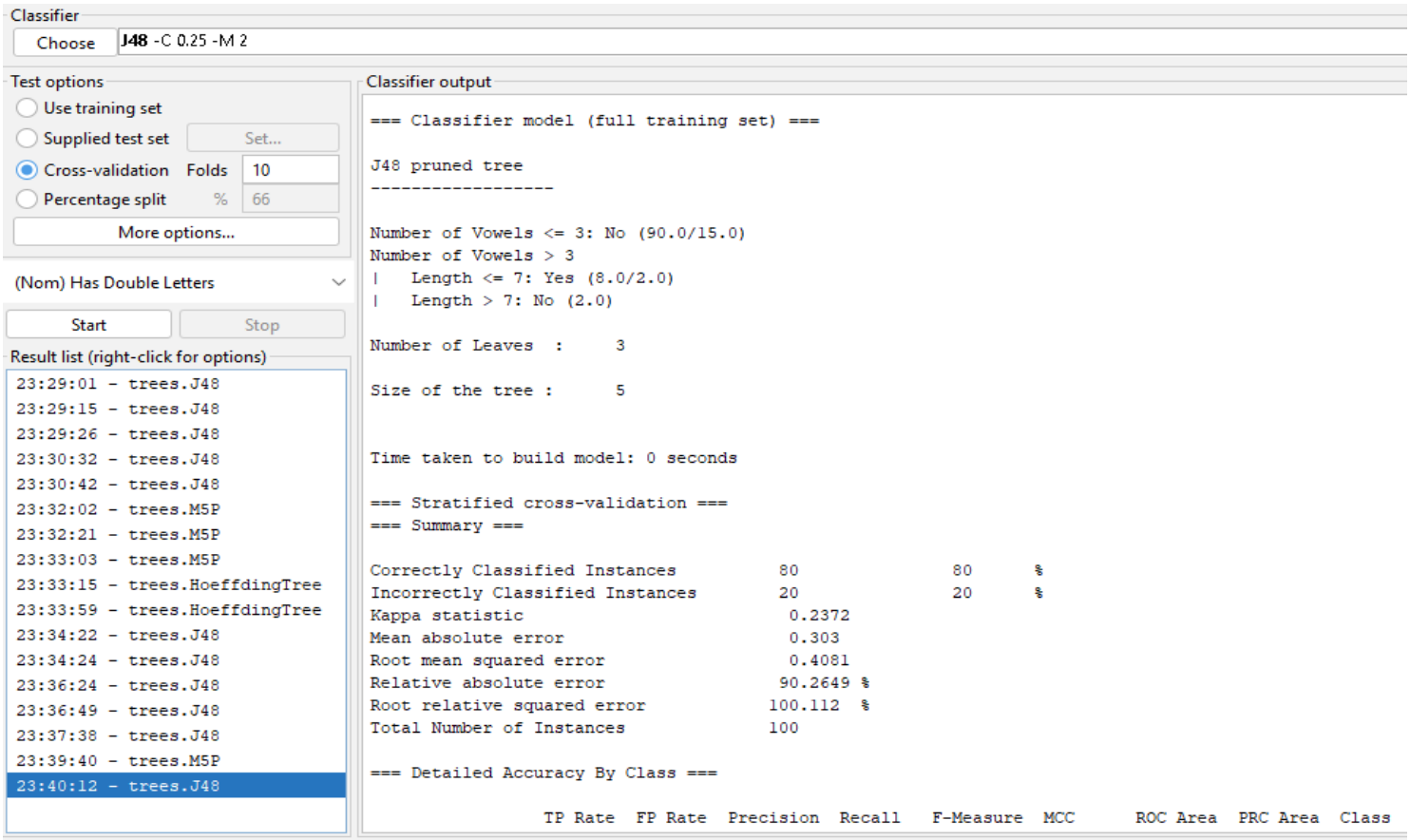


1) **Query: If you notice anything interesting about the dataset, record it.**

The data set has 100 instances and 11 attributes and it showed that the data set has 99 distinct instances and 98% Unique. After observing the data I found out that the name “Abeer” was typed twice with same result/Output. That is the reason it was 98% unique neglecting the 2% where “Abeer” was repeated which wasn’t unique.

2) **Results after Running j48 Classification Algorithm along with a few more:**

a) If root attribute is The name contains Double Letters:
The result shows it has 80 correctly classified instances which is close to perfect but not perfect. This result is a very huge development.

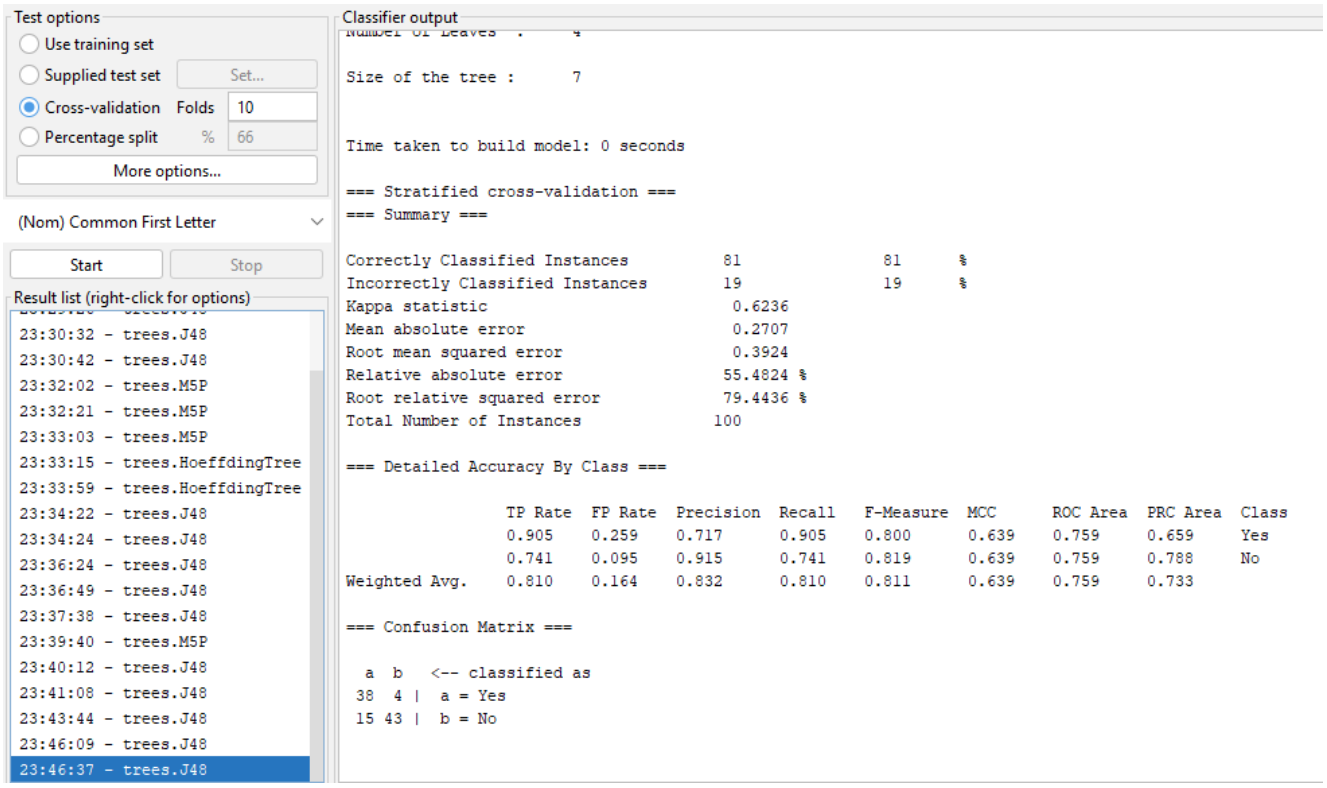



b) If The attribute is Length of the Name:
The result shows it has 81 correctly classified instances which is close to perfect but not perfect. This result is a very huge development.

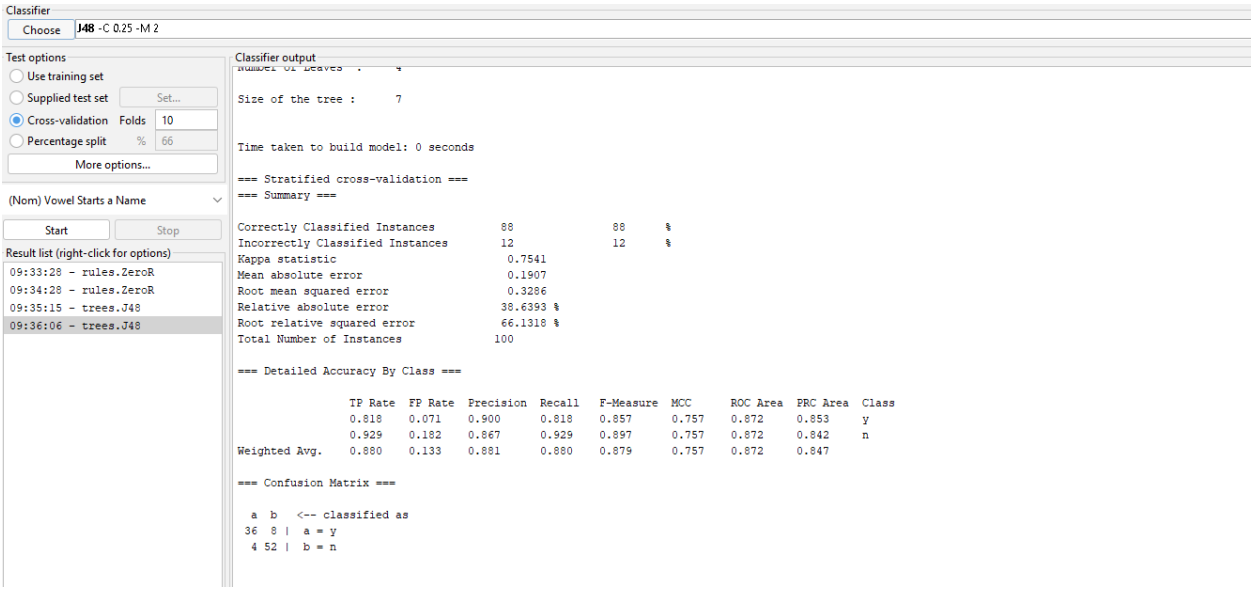
```
|   |   Length > 6: No (5.0/1.0)
|   Second Alphabet is Vowel = Yes: No (45.0/3.0)

Number of Leaves  :     4

Size of the tree :      7


Time taken to build model: 0.02 seconds

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

Correctly Classified Instances          81               81      %
Incorrectly Classified Instances        19               19      %
Kappa statistic                          0.6236
Mean absolute error                      0.2707
Root mean squared error                  0.3924
Relative absolute error                 55.4824 %
Root relative squared error             79.4436 %
Total Number of Instances              100

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.905    0.259    0.717      0.905    0.800      0.639    0.759     0.659     Yes
                 0.741    0.095    0.915      0.741    0.819      0.639    0.759     0.788     No
Weighted Avg.    0.810    0.164    0.832      0.810    0.811      0.639    0.759     0.733

=== Confusion Matrix ===

  a  b   <-- classified as
 38  4 |  a = Yes
 15 43 |  b = No
```

c) For “First Common Letter” :

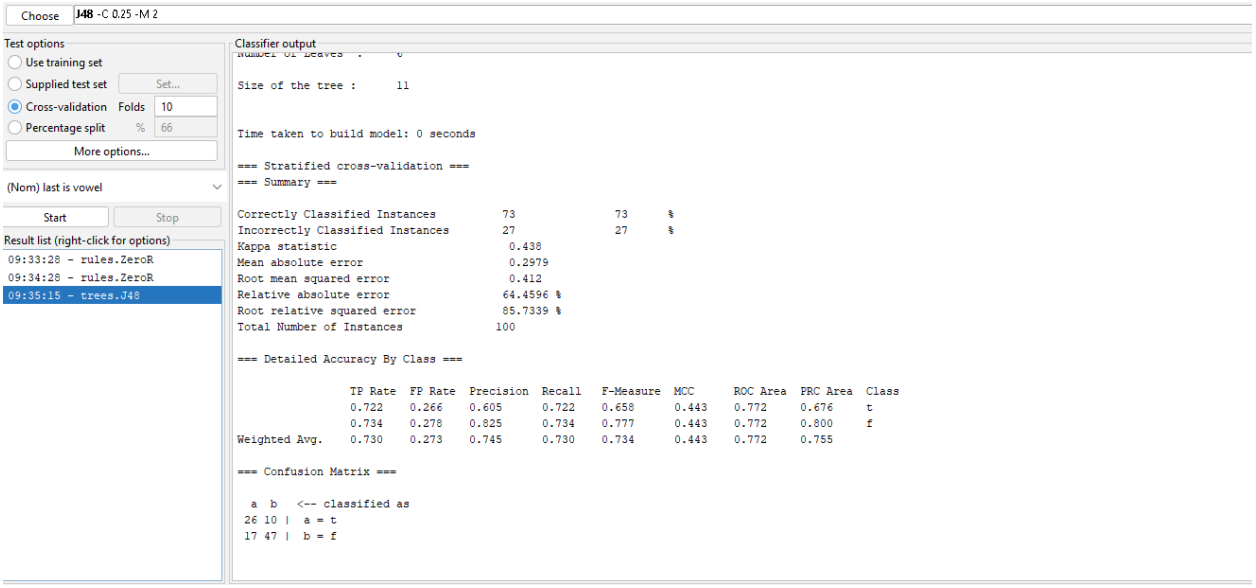
The model achieved 81 correctly classified instances, demonstrating near-perfect performance. While not flawless, this result represents a significant advancement and indicates substantial improvement in the model's accuracy.



d) For First Letter is Vowel:




e) For last Letter is Vowel:




f) For Length of the Name

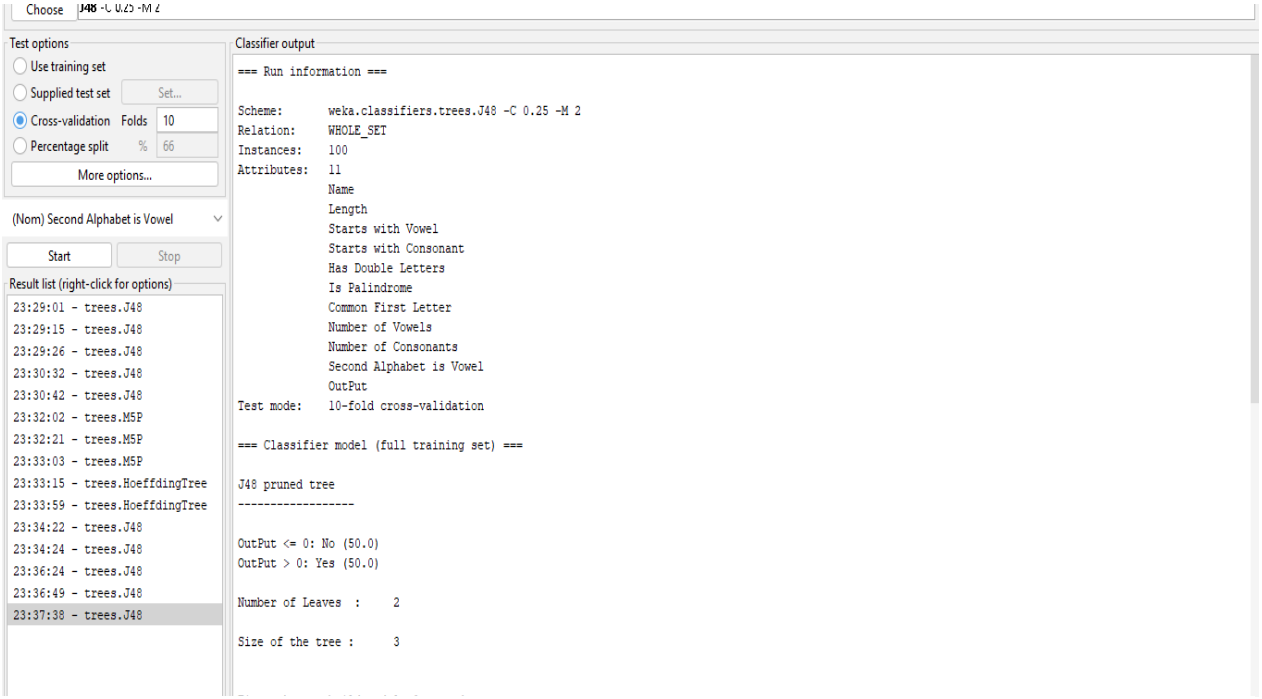
Classifier output

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

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

n (57.000) NB1 NB adaptivel

Time taken to build model: 0.02 seconds

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

Correctly Classified Instances          83               83      %
Incorrectly Classified Instances        17               17      %
Kappa statistic                          0.6592
Mean absolute error                      0.1929
Root mean squared error                  0.3605
Relative absolute error                 39.0827 %
Root relative squared error             72.5453 %
Total Number of Instances              100

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.864    0.196    0.776      0.864    0.817      0.663    0.878     0.861     y
                 0.804    0.136    0.882      0.804    0.841      0.663    0.878     0.846     n
Weighted Avg.    0.830    0.163    0.835      0.830    0.831      0.663    0.878     0.852

=== Confusion Matrix ===

  a  b   <-- classified as
 38  6 |  a = y
 11 45 |  b = n
```

g) For Second Alphabet is vowel or Consonant
For this it shows that correctly classified instances are 100 which is best possible outcome for a classifier. It highlights that this is the magical attribute we are looking for and it points out that:

- Clearly state that this specific attribute has complete predictive power over the target class.
- This attribute alone can be used as a rule-based classifier, achieving 100% accuracy, so it simplifies the classification task since no other attribute or complex model is necessary.

Upon analysis, this attribute perfectly classifies the dataset with 100% accuracy, indicating that it fully determines the class label without error. This attribute can serve as a standalone classifier, making the task of classification straightforward.




```
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
```

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

Working with the standard machine learning pipeline involves several crucial steps, from data preprocessing to model evaluation. For me this process began with feature extraction, where I analyzed a dataset containing names and manually derived relevant features such as

- The length of the name
- Vowel and consonant counts
- Starts with Vowel and consonant
- If name had specific patterns like double letters
- If it started or Ended with a vowel and more

But this did not give me the result I wanted. So I searched, added and tested more features, which at some time gave better results and sometimes the algorithm performed not so well. But at the end I found the magical feature which is "***Second Alphabet is a Vowel***". Given task of manual feature Engineering in this Second Assignment is critical in influencing model performance by providing the algorithm with meaningful insights. Once the dataset was structured with both input and output features, I converted it into ARFF format by creating a label of @relation at the top of file, which WEKA—a machine learning tool—can read and process. The ARFF file was loaded into WEKA for experimentation, where I explored the data's characteristics, such as attribute distributions and class labels. I then ran the J48 decision tree algorithm to classify the data based on the features I had crafted. After observing the performances I added, deleted and altered the features I found fit. This helped me gain deeper and precise understanding of this algorithm. Throughout this process, I gained a deeper understanding of how each stage in the pipeline such as:

- Data preparation
- Feature engineering
- Relation of each attribute with output
- Model training
- Evaluation

plays a role in shaping the effectiveness of a machine learning model.