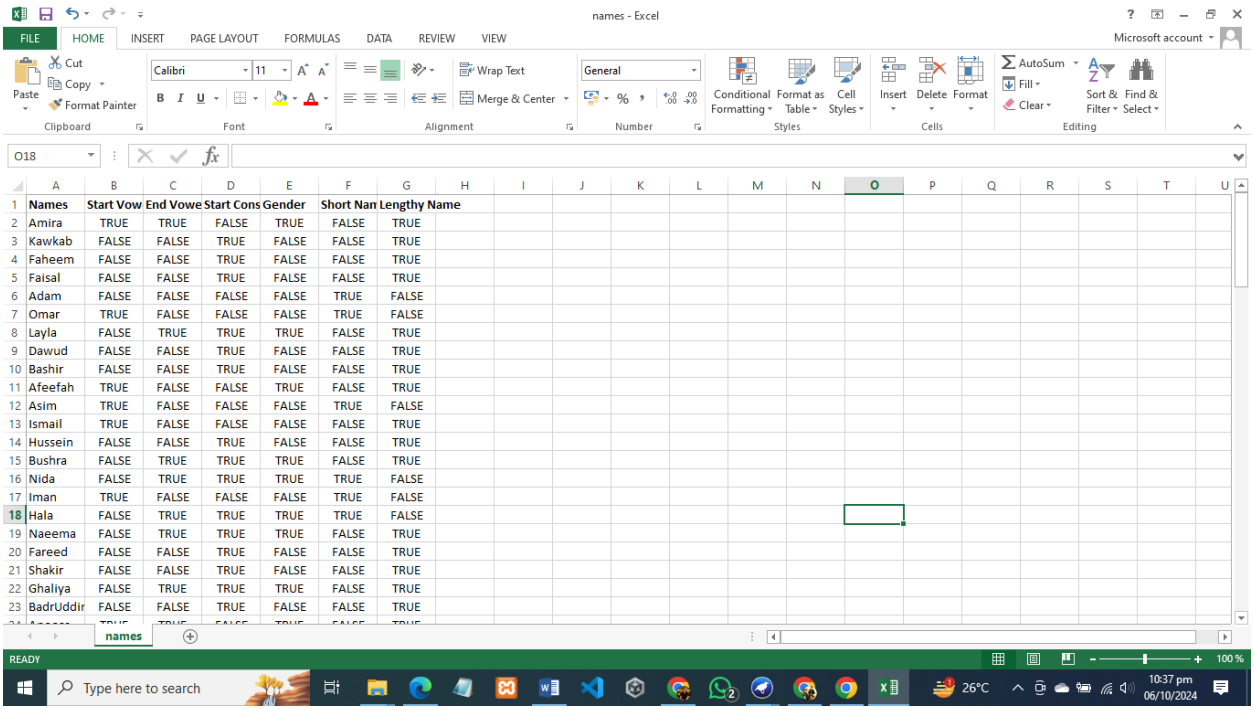
	Course: Machine Learning	**Instructor:** Dr. M Sharjeel
	Name: Muhammad Rabi	**Reg No:** SP24-RCS-007
	Assignment: 02	**Date:** 06/10/2024

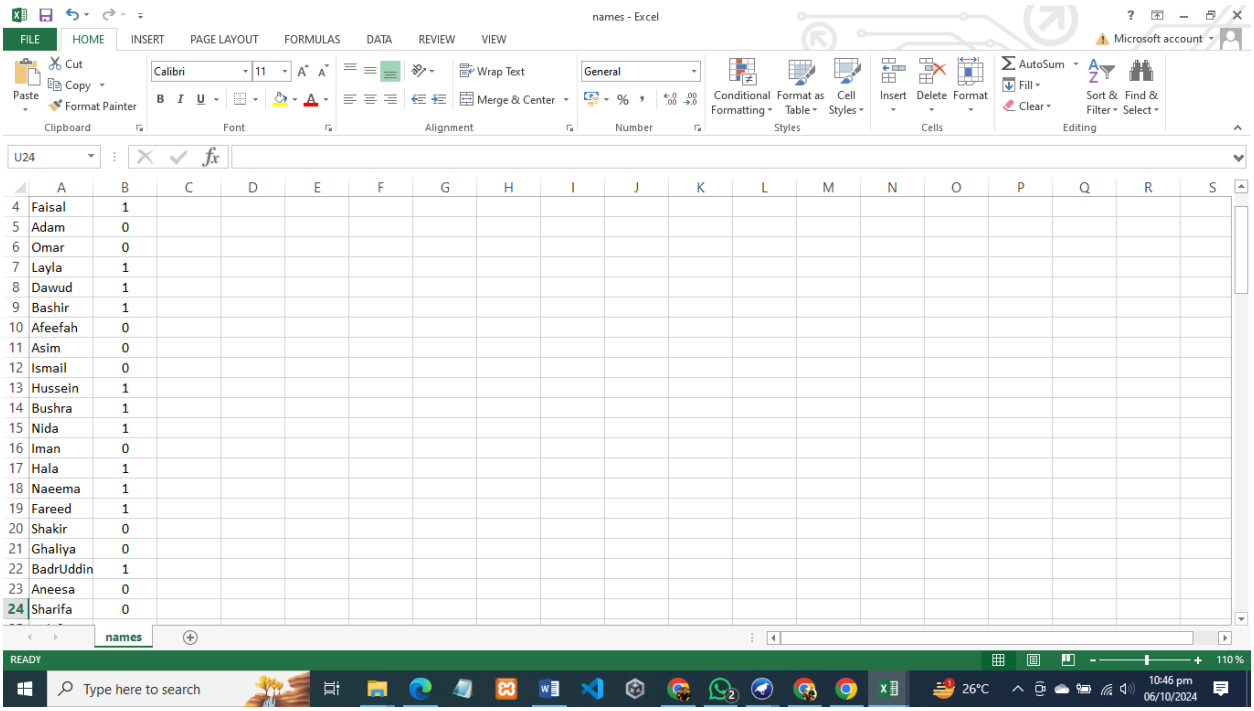
Q:1 Hand crafted features:

Extract as many input features as you can by manually observing the text, i.e., the names of people. Create ARFF file(s). You can save the features together as a set of input features or separately one feature per file. Hint: Remember the input feature(s) is the key to getting a good performance from the classifier



Names	Start Vow	End Vowe	Start Cons	Gender	Short Nan	Lengthy Name
Amira	TRUE	TRUE	FALSE	TRUE	FALSE	TRUE
Kawkab	FALSE	FALSE	TRUE	FALSE	FALSE	TRUE
Faheem	FALSE	FALSE	TRUE	FALSE	FALSE	TRUE
Faisal	FALSE	FALSE	TRUE	FALSE	FALSE	TRUE
Adam	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
Omar	TRUE	FALSE	FALSE	FALSE	TRUE	FALSE
Layla	FALSE	TRUE	TRUE	TRUE	FALSE	TRUE
Dawud	FALSE	FALSE	TRUE	FALSE	FALSE	TRUE
Bashir	FALSE	FALSE	TRUE	FALSE	FALSE	TRUE
Afeefah	TRUE	FALSE	FALSE	TRUE	FALSE	TRUE
Asim	TRUE	FALSE	FALSE	FALSE	TRUE	FALSE
Ismail	TRUE	FALSE	FALSE	FALSE	FALSE	TRUE
Hussein	FALSE	FALSE	TRUE	FALSE	FALSE	TRUE
Bushra	FALSE	TRUE	TRUE	TRUE	FALSE	TRUE
Nida	FALSE	TRUE	TRUE	TRUE	TRUE	FALSE
Iman	TRUE	FALSE	FALSE	FALSE	TRUE	FALSE
Hala	FALSE	TRUE	TRUE	TRUE	TRUE	FALSE
Naeema	FALSE	TRUE	TRUE	TRUE	FALSE	TRUE
Fareed	FALSE	FALSE	TRUE	FALSE	FALSE	TRUE
Shakir	FALSE	FALSE	TRUE	FALSE	FALSE	TRUE
Ghaliya	FALSE	TRUE	TRUE	TRUE	FALSE	TRUE
BadrUddir	FALSE	FALSE	TRUE	FALSE	FALSE	TRUE

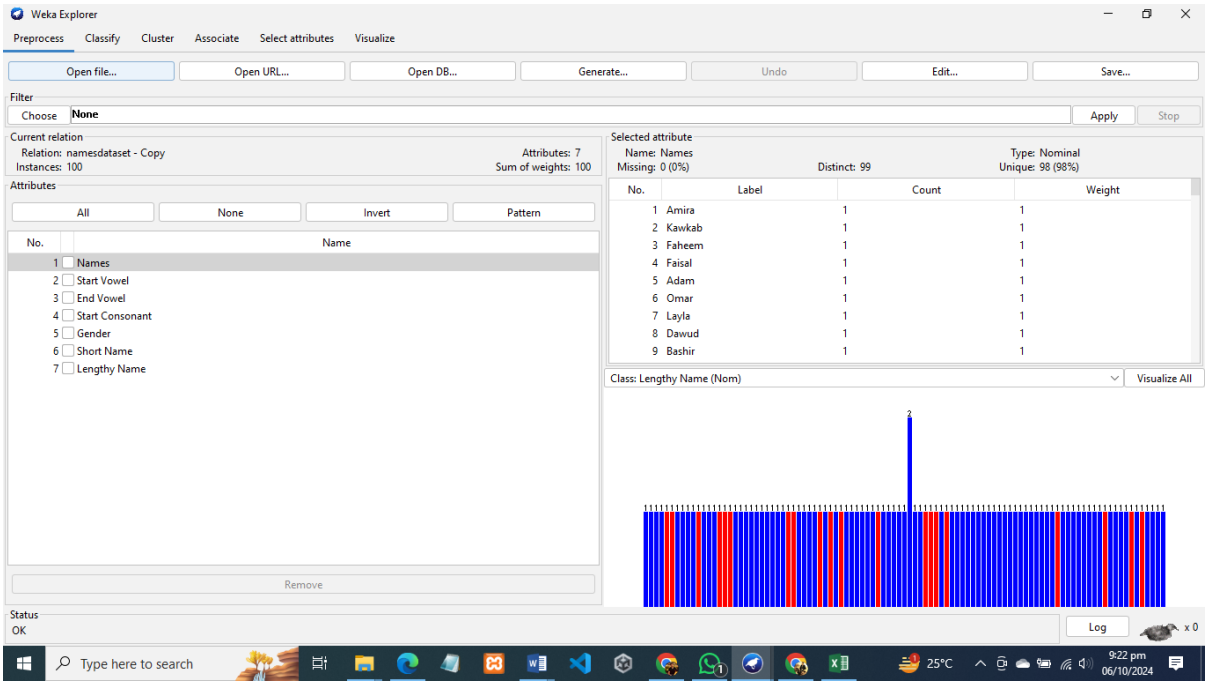
Convert the output feature, i.e., + and – symbols to their numeric equivalent (1 and 0).



	A	B
4	Faisal	1
5	Adam	0
6	Omar	0
7	Layla	1
8	Dawud	1
9	Bashir	1
10	Afeefah	0
11	Asim	0
12	Ismail	0
13	Hussein	1
14	Bushra	1
15	Nida	1
16	Iman	0
17	Hala	1
18	Naeema	1
19	Fareed	1
20	Shakir	0
21	Ghaliya	0
22	BadrUddin	1
23	Aneesa	0
24	Sharifa	0

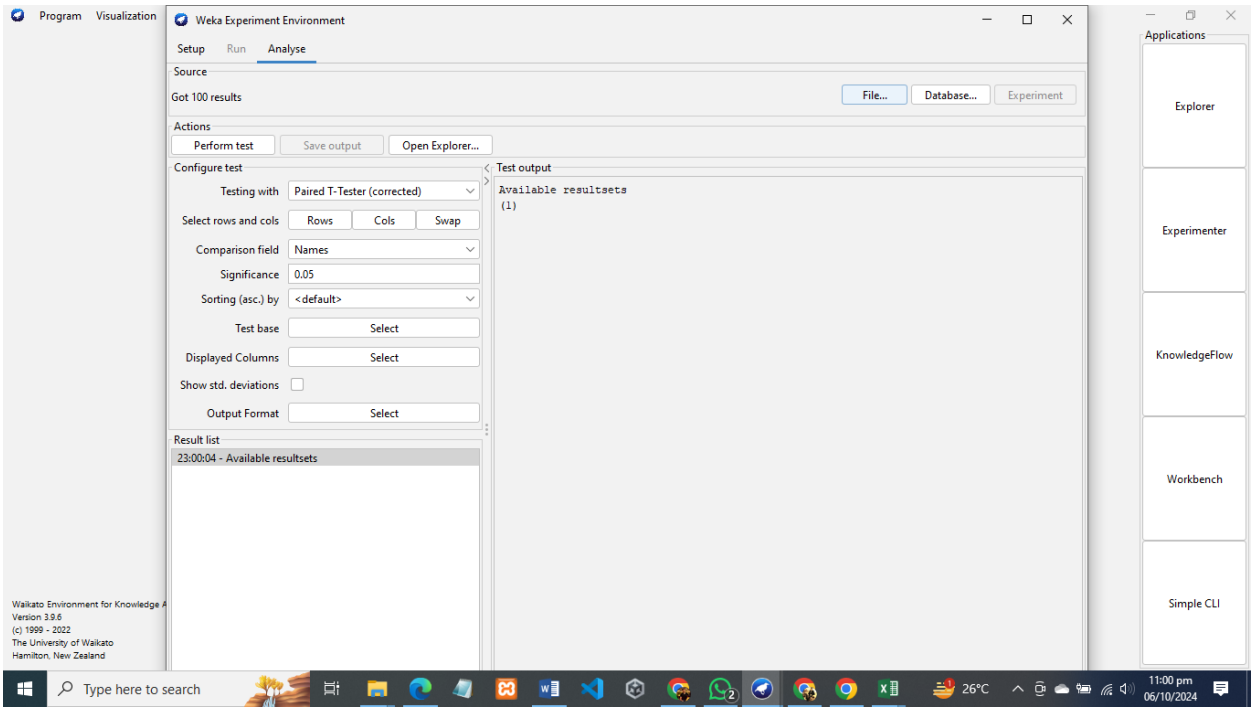
Q:2 ML experiments in WEKA:

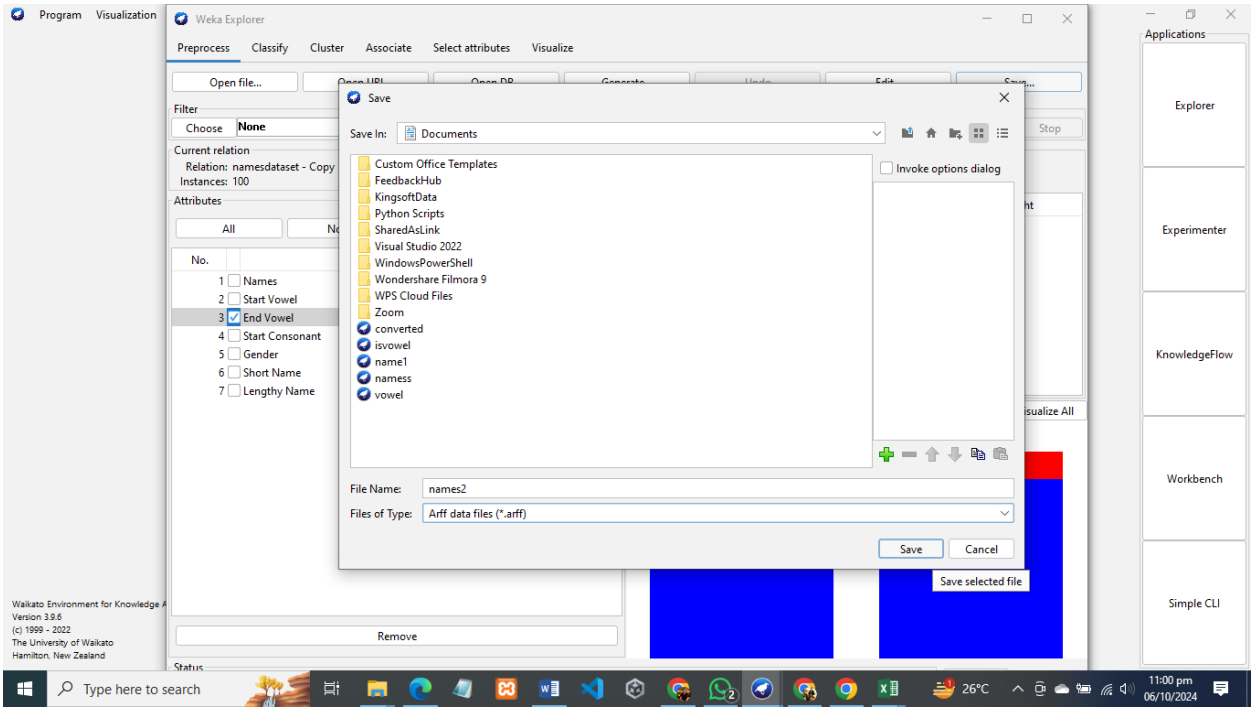
Once you have the ARFF file(s) ready, load it into the WEKA's workbench.



View different characteristics of the data (WEKA's main window). If you notice anything interesting about the dataset, record it.

We can convert our CSV file to ARFF file directly in WEKA.





Run the j48 classification algorithm and observe/record the results.

Weka Explorer

Preprocess | Classify | Cluster | Associate | Select attributes | Visualize

Classifier

Choose **J48** -C 0.25 -M 2

Test options

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

More options...

(Nom) Lengthy Name

Start | Stop

Result list (right-click for options)

21:22:53 - trees.J48

Classifier output

```
=== Run information ===

Scheme:       weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:     namesdataset - Copy
Instances:    100
Attributes:   7
              Names
              Start Vowel
              End Vowel
              Start Consonant
              Gender
              Short Name
              Lengthy Name
Test mode:    10-fold cross-validation

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

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

Short Name = FALSE: TRUE (80.0)
Short Name = TRUE: FALSE (20.0)

Number of Leaves  :     2

Size of the tree :      3


Time taken to build model: 0 seconds

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

Status: OK

Log x 0

Type here to search | 25°C | 9:23 pm 06/10/2024

Weka Explorer

Preprocess | Classify | Cluster | Associate | Select attributes | Visualize

Classifier

Choose **J48** -C 0.25 -M 2

Test options

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

More options...

(Nom) Lengthy Name

Start | Stop

Result list (right-click for options)

21:22:53 - trees.J48

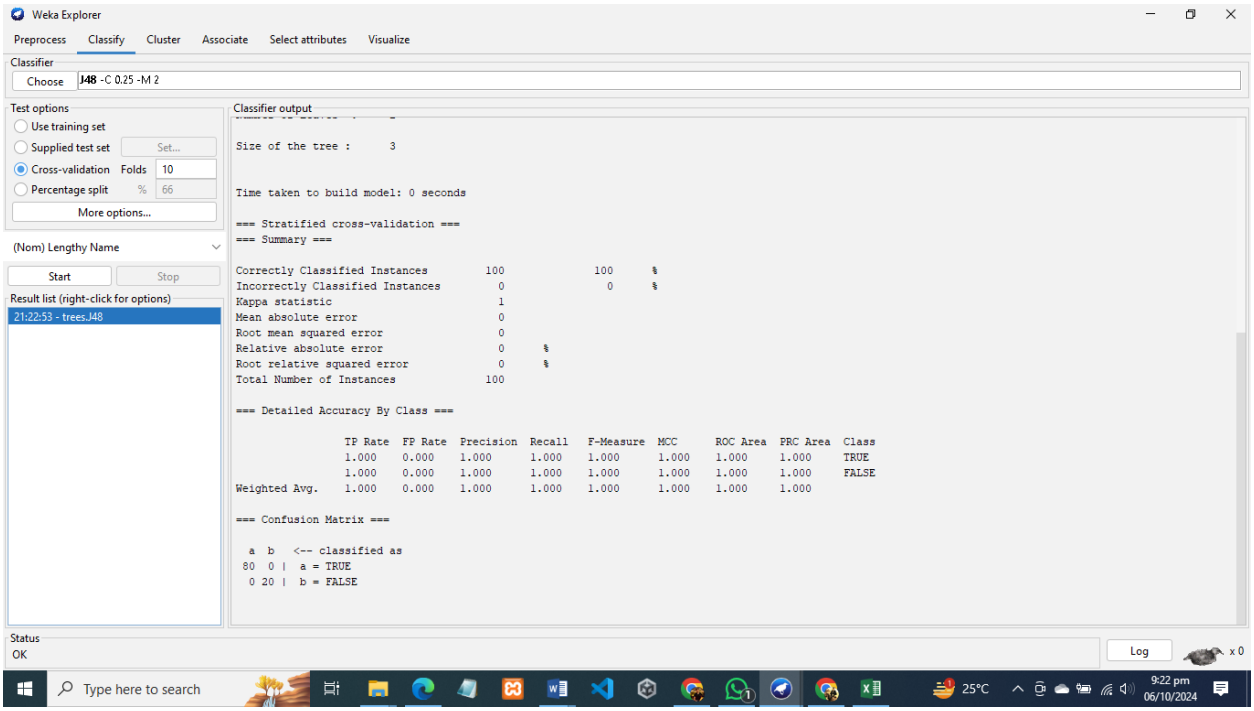
Classifier output

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

Short Name = FALSE: TRUE (80.0)
Short Name = TRUE: FALSE (20.0)

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     TRUE
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     FALSE
Weighted Avg.    1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000
```

Status: OK

Log x 0

Type here to search | 25°C | 9:23 pm 06/10/2024



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

The model performs exceptionally well with 100% accuracy, likely due to the simplicity of the decision tree and the nature of the dataset. It perfectly classifies instances based on the Short Name attribute. However, in real-world applications, such perfect results may indicate that the dataset is either very simple or that there is a risk of overfitting.

1. Scheme:

- **Algorithm**: J48 decision tree algorithm

2. Dataset Information:

- **Dateset Name**: namesdataset.arff

- **Instances**: 100
- **Attributes**: 7 features including Names, Start Vowel, End Vowel, Start Consonant, Gender, Short Name, and Lengthy Name.

3. Classifier Model:

- The decision tree is very simple, with only two leaves and a size of 3.
- The decision rule is based on the Short Name attribute:
- If Short Name = FALSE, the predicted class is TRUE (80 instances).
- If Short Name = TRUE, the predicted class is FALSE (20 instances).

4. Model Performance:

- **Correctly Classified Instances**: 100 (100% accuracy).
- **Incorrectly Classified Instances**: 0 (0% error).
- **Mean Absolute Error**: 0 (indicates perfect predictions).
- **Root Mean Squared Error**: 0 (no error).

5. Detailed Accuracy by Class:

- Both classes (TRUE and FALSE) have a True Positive Rate of 1.000 (100%), meaning the model perfectly classifies both classes.
- Precision, Recall, and F-Measure are all 1.000 for both classes, indicating flawless classification.

- **MCC (Matthews Correlation Coefficient)**: 1.000 (a perfect score, indicating a strong relationship between predictions and actual values).
- **ROC Area** & **PRC Area**: Both 1.000, showing perfect discrimination between classes.

6. Confusion Matrix:

- 80 instances of the class TRUE were correctly classified.
- 20 instances of the class FALSE were correctly classified.
- No instances were misclassified.