# ETL-PIPELINE(COVID)

## Problem Statement

Our problem statement is to analyze the rate of how much covid is spreading and how much the death toll is day wise month wise and country wise

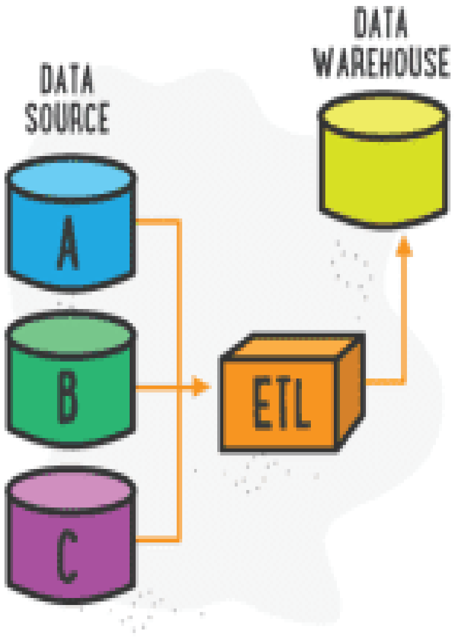
## Proposed Solution

We aim to use a data warehouse mining strategy combined with an Extract, Transform, Load (ETL) process to collect, cleanse, and merge various data sources into a Snowflake schema. This integrated data will then be visualized using Power BI and then we apply weka for clustering

## Expected Outcome

- Improved Decision Making
- Comprehensive Data Analysis
- Scalability and Flexibility

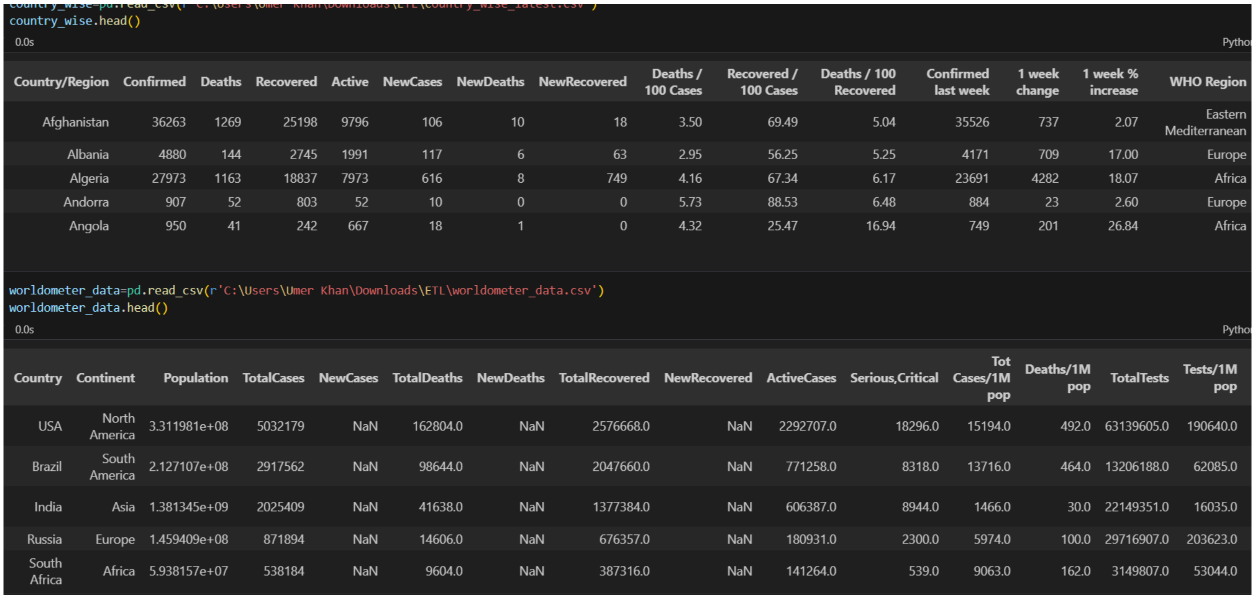
## ETL IMPLEMENTATION



ILLUSTRATED BY SEGUE TECHNOLOGIES

## EXTRACT:

We extract data into Kaggle like country wise ,day wise and world wise and put it into notebook and convert our data into csv format.



```
country_wise=pd.read_csv(r'C:\Users\Umer Khan\Downloads\ETL\country_wise_latest.csv')
country_wise.head()
```

0.0s Pytho

| Country/Region | Confirmed | Deaths | Recovered | Active | NewCases | NewDeaths | NewRecovered | Deaths / 100 Cases | Recovered / 100 Cases | Deaths / 100 Recovered | Confirmed last week | 1 week change | 1 week % increase | WHO Region |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Afghanistan | 36263 | 1269 | 25198 | 9796 | 106 | 10 | 18 | 3.50 | 69.49 | 5.04 | 35526 | 737 | 2.07 | Eastern Mediterranean |
| Albania | 4880 | 144 | 2745 | 1991 | 117 | 6 | 63 | 2.95 | 56.25 | 5.25 | 4171 | 709 | 17.00 | Europe |
| Algeria | 27973 | 1163 | 18837 | 7973 | 616 | 8 | 749 | 4.16 | 67.34 | 6.17 | 23691 | 4282 | 18.07 | Africa |
| Andorra | 907 | 52 | 803 | 52 | 10 | 0 | 0 | 5.73 | 88.53 | 6.48 | 884 | 23 | 2.60 | Europe |
| Angola | 950 | 41 | 242 | 667 | 18 | 1 | 0 | 4.32 | 25.47 | 16.94 | 749 | 201 | 26.84 | Africa |

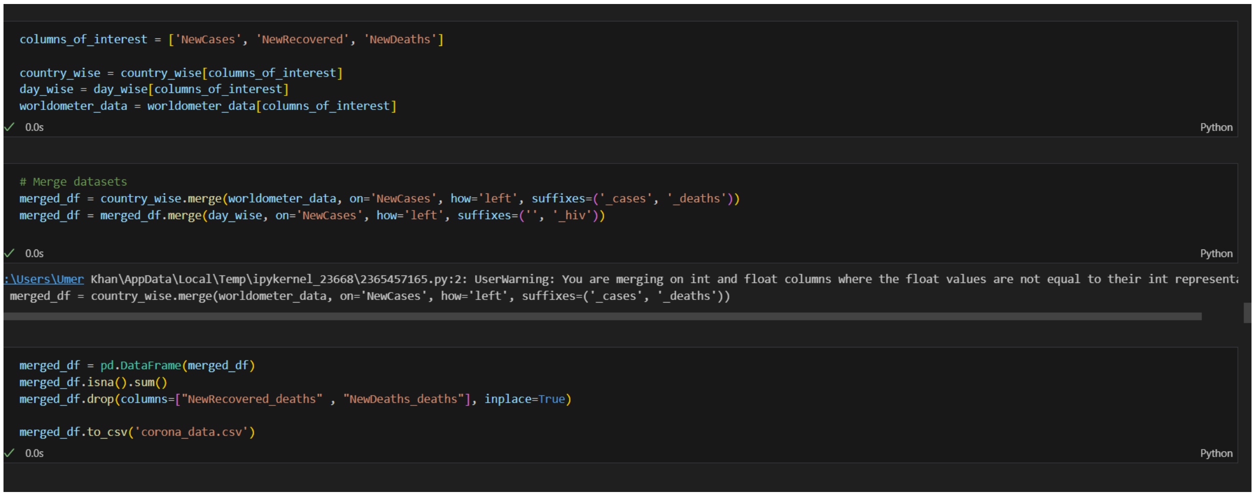
```
worldometer_data=pd.read_csv(r'C:\Users\Umer Khan\Downloads\ETL\worldometer_data.csv')
worldometer_data.head()
```

0.0s Pytho

| Country | Continent | Population | TotalCases | NewCases | TotalDeaths | NewDeaths | TotalRecovered | NewRecovered | ActiveCases | Serious,Critical | Tot Cases/1M pop | Deaths/1M pop | TotalTests | Tests/1M pop |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| USA | North America | 3.311981e+08 | 5032179 | NaN | 162804.0 | NaN | 2576668.0 | NaN | 2292707.0 | 18296.0 | 15194.0 | 492.0 | 63139605.0 | 190640.0 |
| Brazil | South America | 2.127107e+08 | 2917562 | NaN | 98644.0 | NaN | 2047660.0 | NaN | 771258.0 | 8318.0 | 13716.0 | 464.0 | 13206188.0 | 62085.0 |
| India | Asia | 1.381345e+09 | 2025409 | NaN | 41638.0 | NaN | 1377384.0 | NaN | 606387.0 | 8944.0 | 1466.0 | 30.0 | 22149351.0 | 16035.0 |
| Russia | Europe | 1.459409e+08 | 871894 | NaN | 14606.0 | NaN | 676357.0 | NaN | 180931.0 | 2300.0 | 5974.0 | 100.0 | 29716907.0 | 203623.0 |
| South Africa | Africa | 5.938157e+07 | 538184 | NaN | 9604.0 | NaN | 387316.0 | NaN | 141264.0 | 539.0 | 9063.0 | 162.0 | 3149807.0 | 53044.0 |

## TRANSFORM/CLEAN:

Then we transform and clean that data and use drop() for dropping null values columns and rows and convert and summarize our data and then merge all sources into one csv file.



```python
columns_of_interest = ['NewCases', 'NewRecovered', 'NewDeaths']

country_wise = country_wise[columns_of_interest]
day_wise = day_wise[columns_of_interest]
worldometer_data = worldometer_data[columns_of_interest]
```

✓ 0.0s Python

```python
# Merge datasets
merged_df = country_wise.merge(worldometer_data, on='NewCases', how='left', suffixes=('_cases', '_deaths'))
merged_df = merged_df.merge(day_wise, on='NewCases', how='left', suffixes=('', '_hiv'))
```

✓ 0.0s Python

```
:\Users\Umer Khan\AppData\Local\Temp\ipykernel_23668\2365457165.py:2: UserWarning: You are merging on int and float columns where the float values are not equal to their int representa
 merged_df = country_wise.merge(worldometer_data, on='NewCases', how='left', suffixes=('_cases', '_deaths'))
```

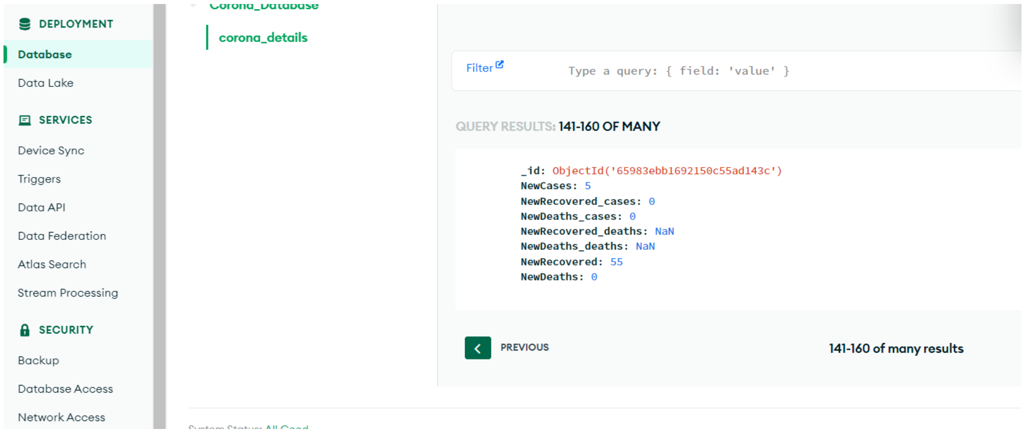
```python
merged_df = pd.DataFrame(merged_df)
merged_df.isna().sum()
merged_df.drop(columns=["NewRecovered_deaths" , "NewDeaths_deaths"], inplace=True)

merged_df.to_csv('corona_data.csv')
```

✓ 0.0s Python

# LOAD:

Then we load our data into mongodb .

I show you the sample:

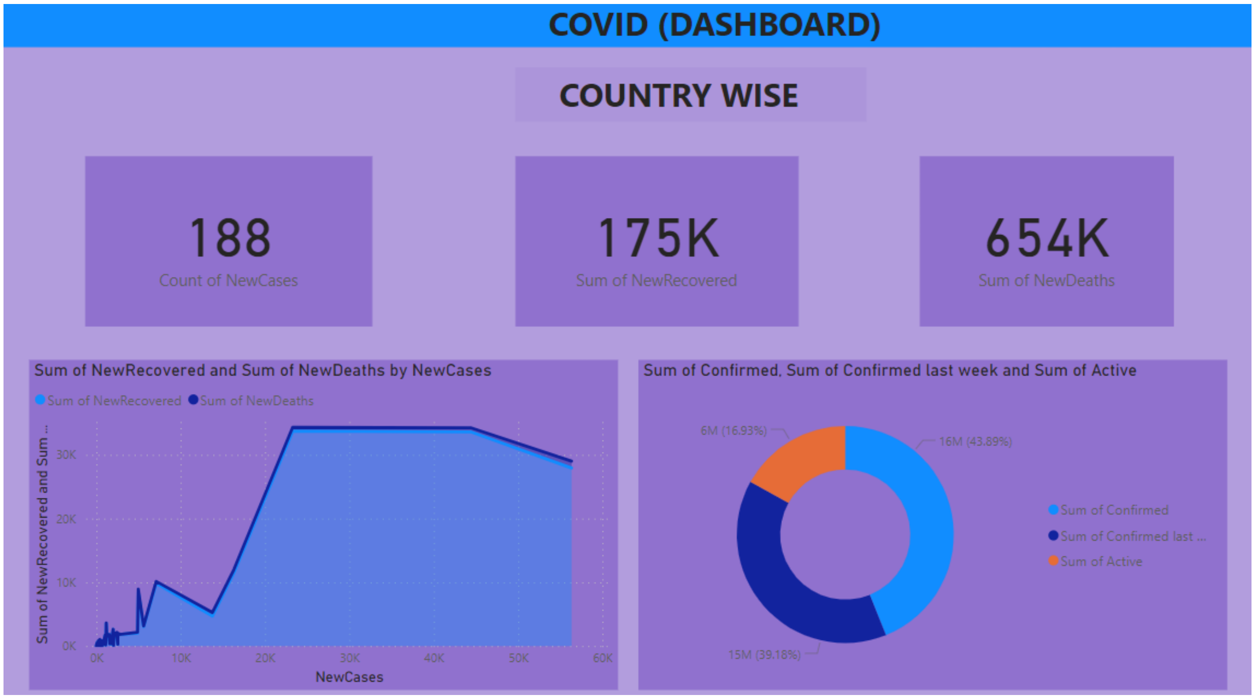


QUERY RESULTS: 141-160 OF MANY

```
_id: ObjectId('65983ebb1692150c55ad143c')
NewCases: 5
NewRecovered_cases: 0
NewDeaths_cases: 0
NewRecovered_deaths: NaN
NewDeaths_deaths: NaN
NewRecovered: 55
NewDeaths: 0
```

141-160 of many results

# POWER BI:



COVID (DASHBOARD)

COUNTRY WISE

188
Count of NewCases

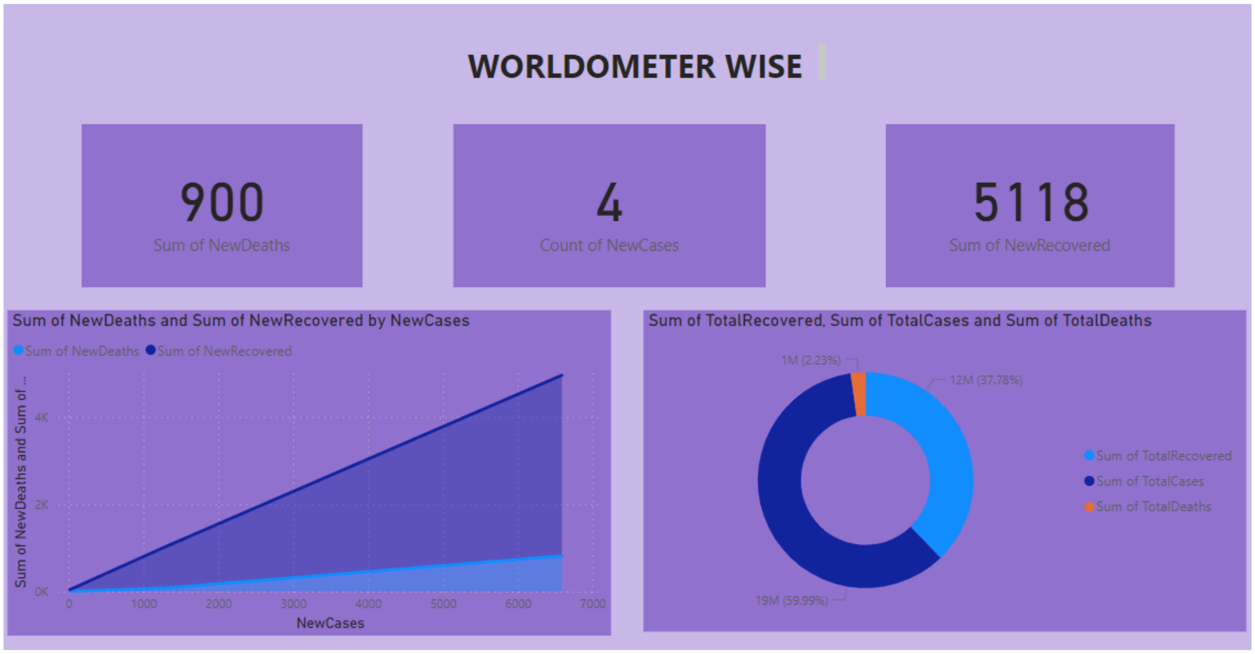
175K
Sum of NewRecovered

654K
Sum of NewDeaths

Sum of NewRecovered and Sum of NewDeaths by NewCases

● Sum of NewRecovered ● Sum of NewDeaths
Sum of NewRecovered and Sum ...
30K
20K
10K
0K
0K
10K
20K
30K
40K
50K
60K
NewCases

Sum of Confirmed, Sum of Confirmed last week and Sum of Active



WORLDOMETER WISE

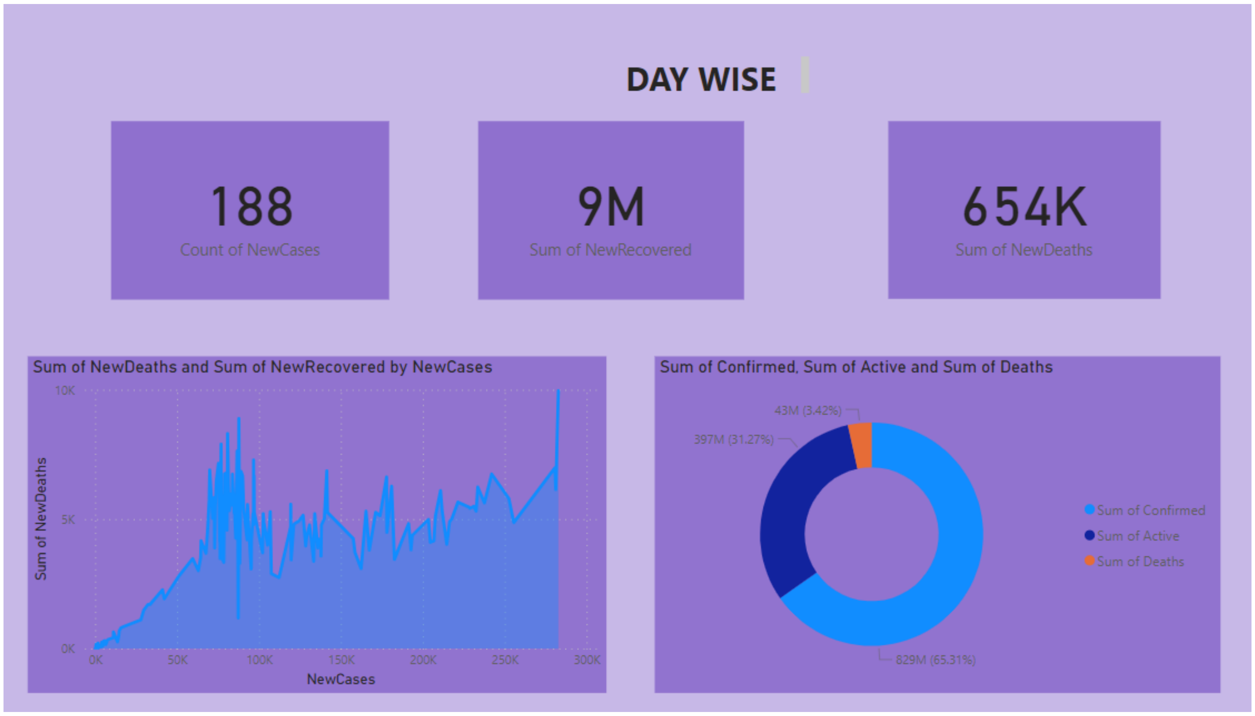
900
Sum of NewDeaths

4
Count of NewCases

5118
Sum of NewRecovered

Sum of NewDeaths and Sum of NewRecovered by NewCases

Sum of TotalRecovered, Sum of TotalCases and Sum of TotalDeaths



DAY WISE

188
Count of NewCases

9M
Sum of NewRecovered

654K
Sum of NewDeaths

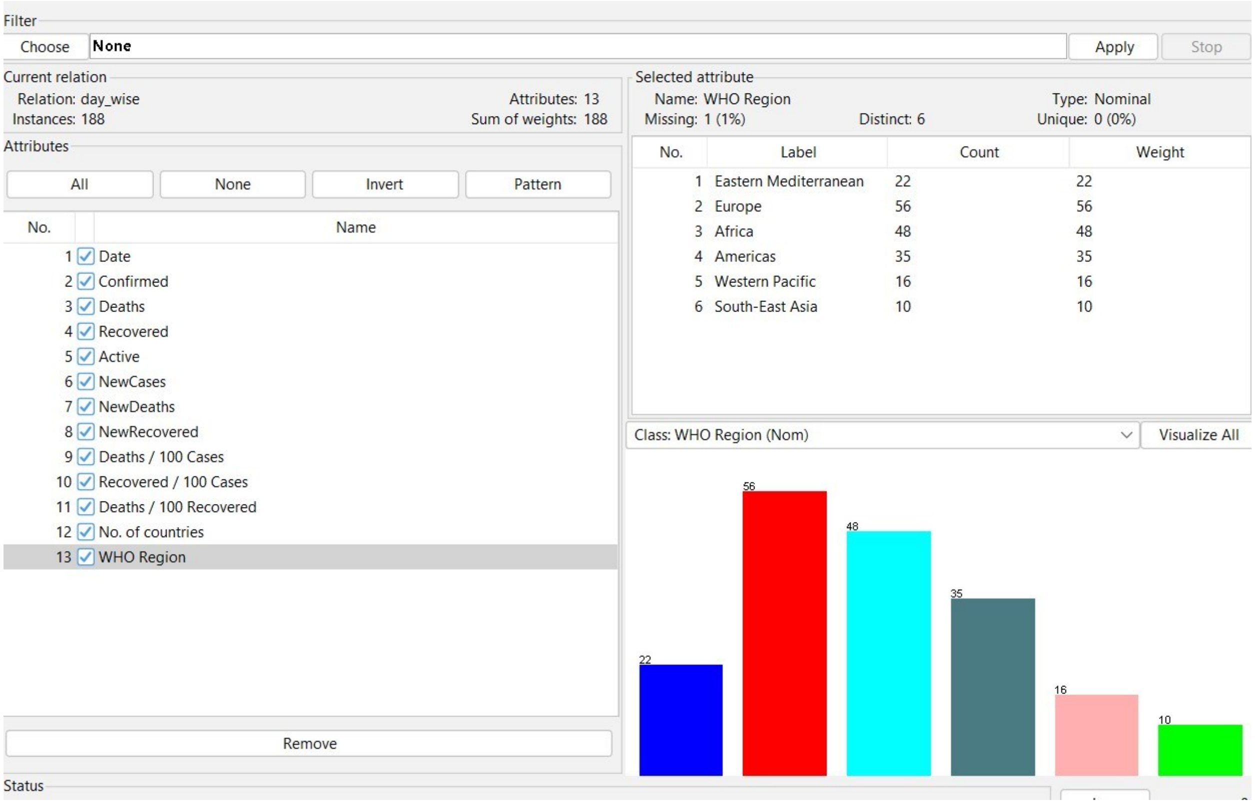
Sum of NewDeaths and Sum of NewRecovered by NewCases

Sum of Confirmed, Sum of Active and Sum of Deaths

# WEKA

## DAY_WISE

### LOAD



Filter: None

Current relation
Relation: day_wise
Attributes: 13
Instances: 188
Sum of weights: 188

Attributes

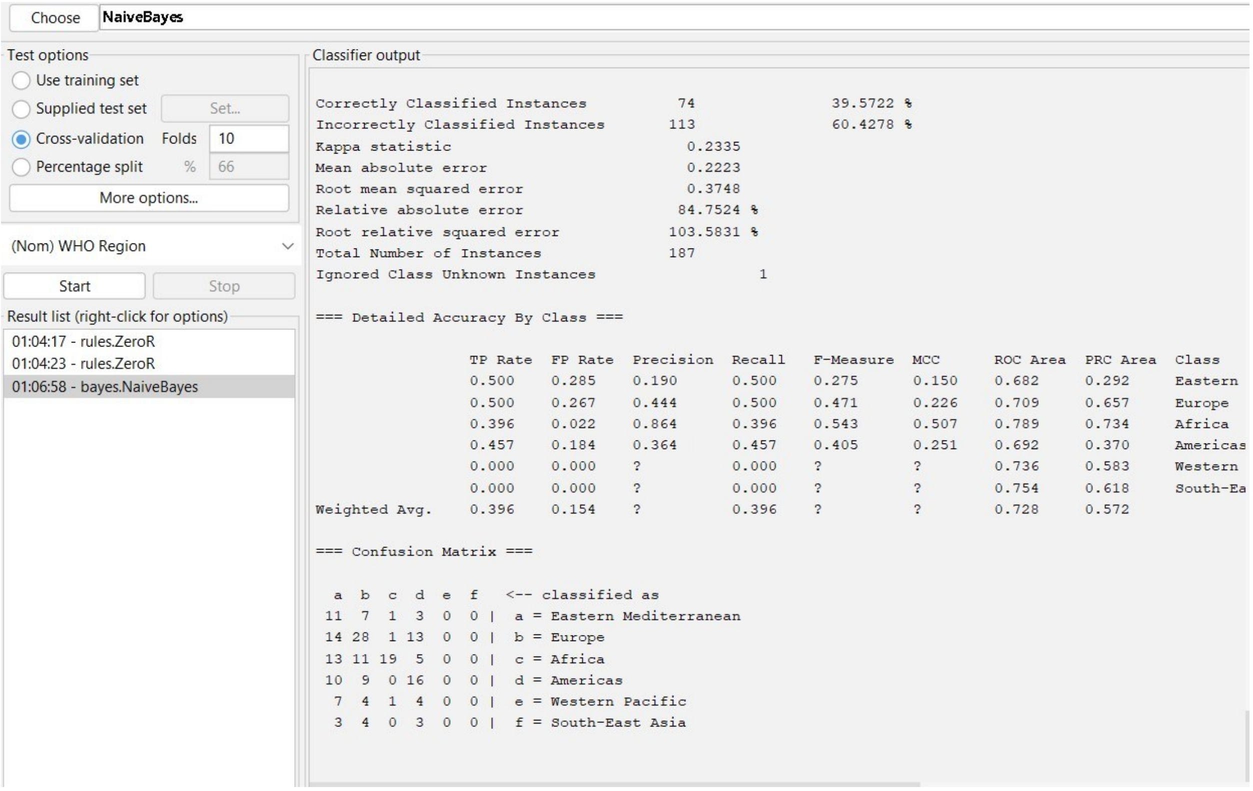
| No. | Name |
|---|---|
| 1 | ☑ Date |
| 2 | ☑ Confirmed |
| 3 | ☑ Deaths |
| 4 | ☑ Recovered |
| 5 | ☑ Active |
| 6 | ☑ NewCases |
| 7 | ☑ NewDeaths |
| 8 | ☑ NewRecovered |
| 9 | ☑ Deaths / 100 Cases |
| 10 | ☑ Recovered / 100 Cases |
| 11 | ☑ Deaths / 100 Recovered |
| 12 | ☑ No. of countries |
| 13 | ☑ WHO Region |

Selected attribute
Name: WHO Region
Type: Nominal
Missing: 1 (1%)
Distinct: 6
Unique: 0 (0%)

| No. | Label | Count | Weight |
|---|---|---|---|
| 1 | Eastern Mediterranean | 22 | 22 |
| 2 | Europe | 56 | 56 |
| 3 | Africa | 48 | 48 |
| 4 | Americas | 35 | 35 |
| 5 | Western Pacific | 16 | 16 |
| 6 | South-East Asia | 10 | 10 |

Class: WHO Region (Nom)

# CLASSIFICATION



NaiveBayes

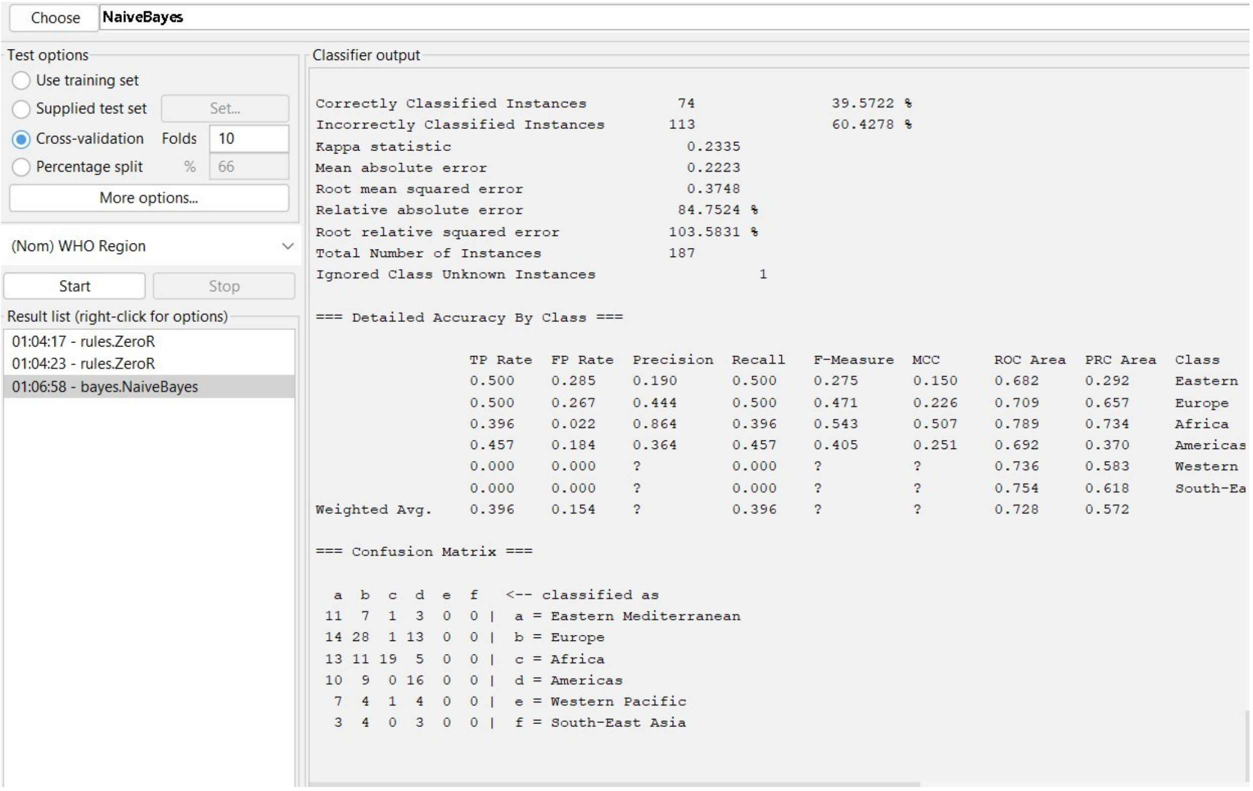
○ Use training set
○ Supplied test set Set...
● Cross-validation Folds 10
○ Percentage split % 66

(Nom) WHO Region

01:04:17 - rules.ZeroR
01:04:23 - rules.ZeroR
01:06:58 - bayes.NaiveBayes

```
Correctly Classified Instances          74               39.5722 %
Incorrectly Classified Instances       113               60.4278 %
Kappa statistic                          0.2335
Mean absolute error                      0.2223
Root mean squared error                  0.3748
Relative absolute error                 84.7524 %
Root relative squared error            103.5831 %
Total Number of Instances              187
Ignored Class Unknown Instances                  1

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.500    0.285    0.190      0.500    0.275      0.150    0.682     0.292     Eastern
                 0.500    0.267    0.444      0.500    0.471      0.226    0.709     0.657     Europe
                 0.396    0.022    0.864      0.396    0.543      0.507    0.789     0.734     Africa
                 0.457    0.184    0.364      0.457    0.405      0.251    0.692     0.370     Americas
                 0.000    0.000    ?          0.000    ?          ?        0.736     0.583     Western
                 0.000    0.000    ?          0.000    ?          ?        0.754     0.618     South-Ea
Weighted Avg.    0.396    0.154    ?          0.396    ?          ?        0.728     0.572

=== Confusion Matrix ===

  a  b  c  d  e  f   <-- classified as
 11  7  1  3  0  0 |  a = Eastern Mediterranean
 14 28  1 13  0  0 |  b = Europe
 13 11 19  5  0  0 |  c = Africa
 10  9  0 16  0  0 |  d = Americas
  7  4  1  4  0  0 |  e = Western Pacific
  3  4  0  3  0  0 |  f = South-East Asia
```

# CLUSTERING



NaiveBayes

Test options

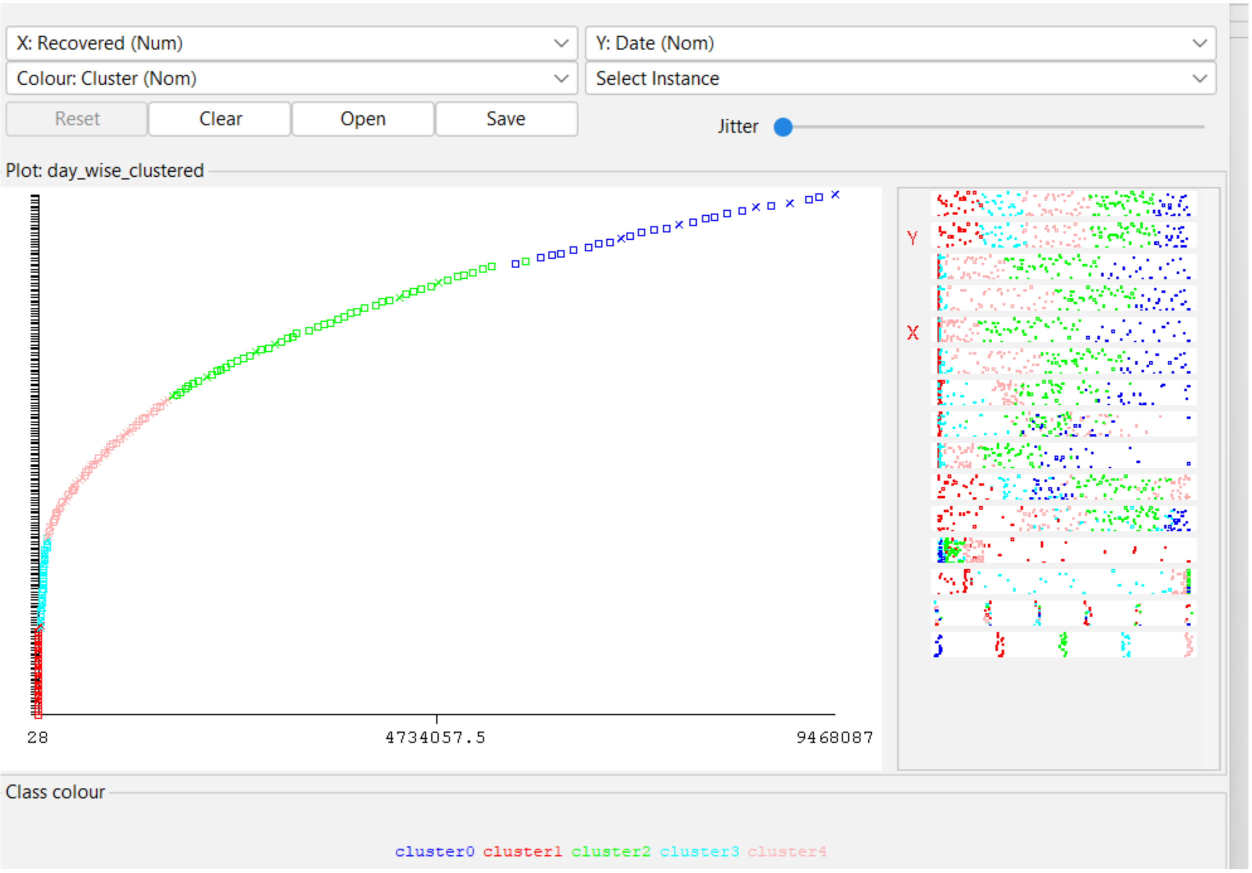
○ Use training set
○ Supplied test set
● Cross-validation Folds 10
○ Percentage split % 66

(Nom) WHO Region

Result list (right-click for options)

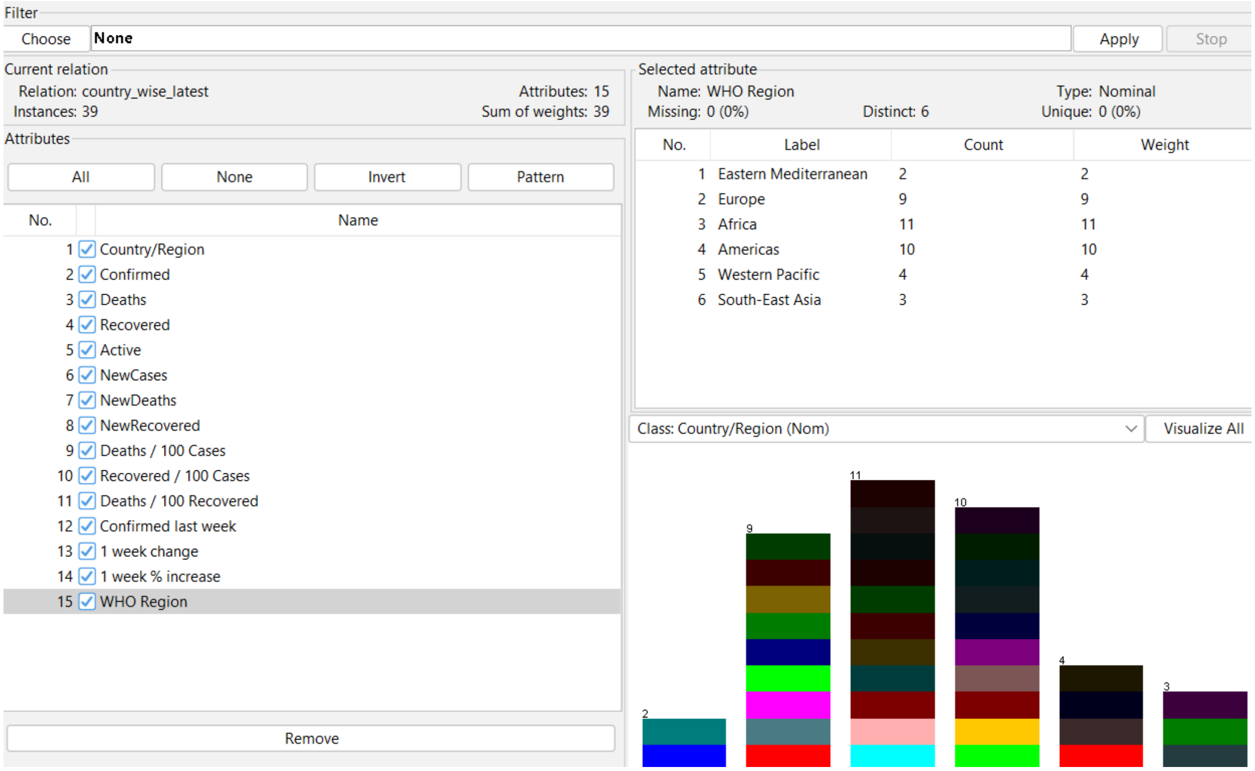
01:04:17 - rules.ZeroR
01:04:23 - rules.ZeroR
01:06:58 - bayes.NaiveBayes

```
Correctly Classified Instances          74               39.5722 %
Incorrectly Classified Instances       113               60.4278 %
Kappa statistic                          0.2335
Mean absolute error                      0.2223
Root mean squared error                  0.3748
Relative absolute error                 84.7524 %
Root relative squared error            103.5831 %
Total Number of Instances              187
Ignored Class Unknown Instances                  1

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.500    0.285    0.190      0.500    0.275      0.150    0.682     0.292     Eastern
                 0.500    0.267    0.444      0.500    0.471      0.226    0.709     0.657     Europe
                 0.396    0.022    0.864      0.396    0.543      0.507    0.789     0.734     Africa
                 0.457    0.184    0.364      0.457    0.405      0.251    0.692     0.370     Americas
                 0.000    0.000    ?          0.000    ?          ?        0.736     0.583     Western
                 0.000    0.000    ?          0.000    ?          ?        0.754     0.618     South-Ea
Weighted Avg.    0.396    0.154    ?          0.396    ?          ?        0.728     0.572

=== Confusion Matrix ===

  a  b  c  d  e  f   <-- classified as
 11  7  1  3  0  0 |  a = Eastern Mediterranean
 14 28  1 13  0  0 |  b = Europe
 13 11 19  5  0  0 |  c = Africa
 10  9  0 16  0  0 |  d = Americas
  7  4  1  4  0  0 |  e = Western Pacific
  3  4  0  3  0  0 |  f = South-East Asia
```



# COUNTRY-WISE

## LOAD



Filter

Choose None Apply Stop

Current relation

Relation: country_wise_latest Attributes: 15

Instances: 39 Sum of weights: 39

Attributes

All None Invert Pattern

| No. | | Name |
|---|---|---|
| 1 | ☑ | Country/Region |
| 2 | ☑ | Confirmed |
| 3 | ☑ | Deaths |
| 4 | ☑ | Recovered |
| 5 | ☑ | Active |
| 6 | ☑ | NewCases |
| 7 | ☑ | NewDeaths |
| 8 | ☑ | NewRecovered |
| 9 | ☑ | Deaths / 100 Cases |
| 10 | ☑ | Recovered / 100 Cases |
| 11 | ☑ | Deaths / 100 Recovered |
| 12 | ☑ | Confirmed last week |
| 13 | ☑ | 1 week change |
| 14 | ☑ | 1 week % increase |
| 15 | ☑ | WHO Region |

Remove

Selected attribute

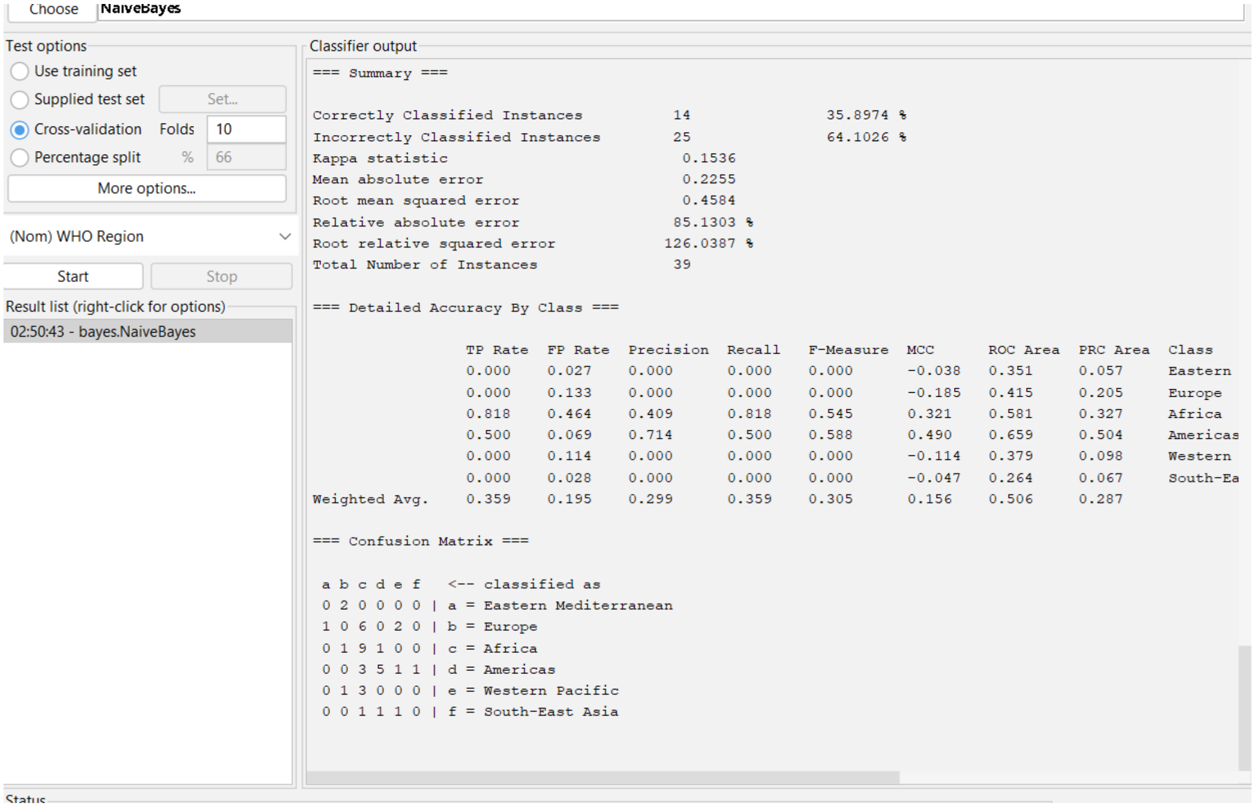
Name: WHO Region Type: Nominal

Missing: 0 (0%) Distinct: 6 Unique: 0 (0%)

| No. | Label | Count | Weight |
|---|---|---|---|
| 1 | Eastern Mediterranean | 2 | 2 |
| 2 | Europe | 9 | 9 |
| 3 | Africa | 11 | 11 |
| 4 | Americas | 10 | 10 |
| 5 | Western Pacific | 4 | 4 |
| 6 | South-East Asia | 3 | 3 |

Class: Country/Region (Nom) Visualize All

# CLASSIFICATION



NaiveBayes

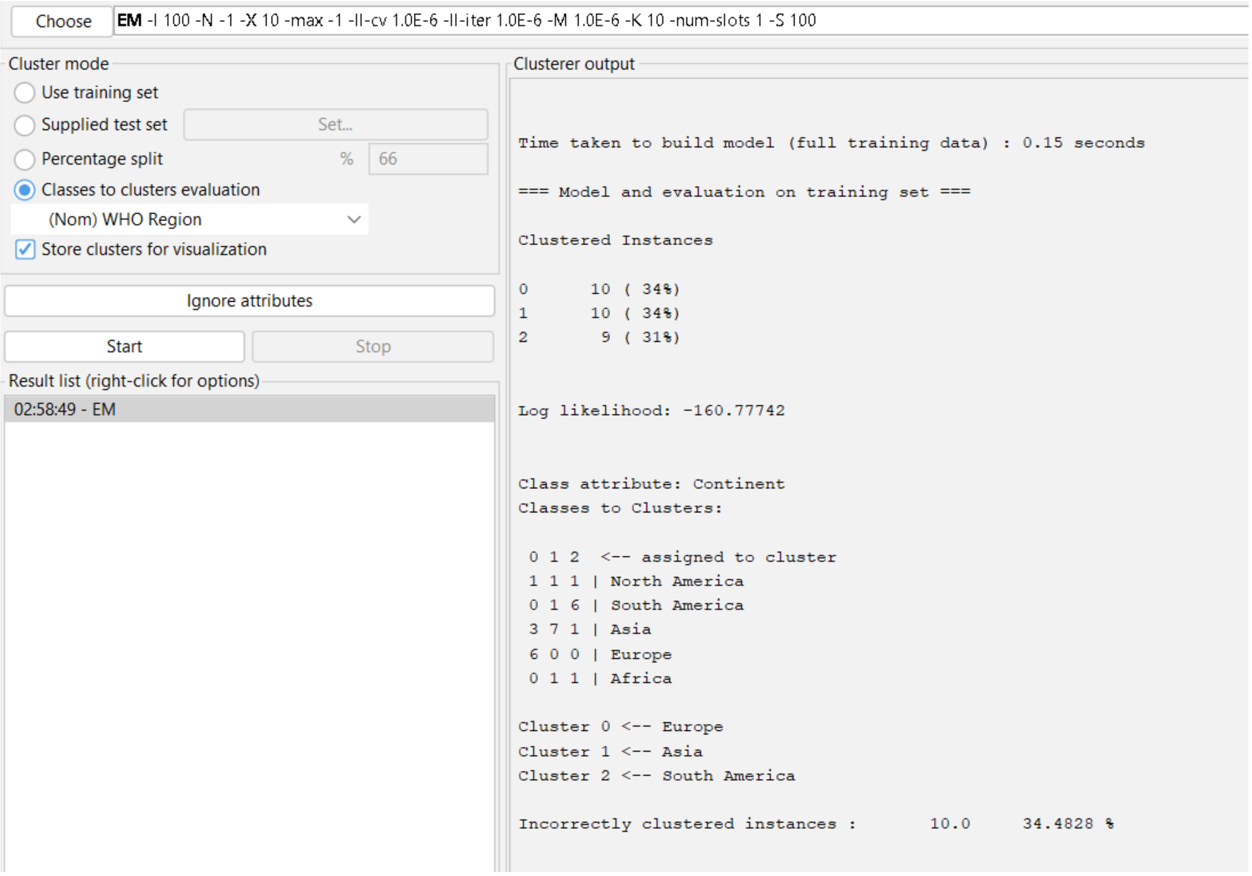
○ Use training set
○ Supplied test set
● Cross-validation Folds 10
○ Percentage split % 66

(Nom) WHO Region

02:50:43 - bayes.NaiveBayes

```
=== Summary ===

Correctly Classified Instances          14               35.8974 %
Incorrectly Classified Instances        25               64.1026 %
Kappa statistic                          0.1536
Mean absolute error                      0.2255
Root mean squared error                  0.4584
Relative absolute error                 85.1303 %
Root relative squared error            126.0387 %
Total Number of Instances               39

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.000    0.027    0.000      0.000    0.000      -0.038   0.351     0.057     Eastern
                 0.000    0.133    0.000      0.000    0.000      -0.185   0.415     0.205     Europe
                 0.818    0.464    0.409      0.818    0.545      0.321    0.581     0.327     Africa
                 0.500    0.069    0.714      0.500    0.588      0.490    0.659     0.504     Americas
                 0.000    0.114    0.000      0.000    0.000      -0.114   0.379     0.098     Western
                 0.000    0.028    0.000      0.000    0.000      -0.047   0.264     0.067     South-Ea
Weighted Avg.    0.359    0.195    0.299      0.359    0.305      0.156    0.506     0.287

=== Confusion Matrix ===

 a b c d e f   <-- classified as
 0 2 0 0 0 0 | a = Eastern Mediterranean
 1 0 6 0 2 0 | b = Europe
 0 1 9 1 0 0 | c = Africa
 0 0 3 5 1 1 | d = Americas
 0 1 3 0 0 0 | e = Western Pacific
 0 0 1 1 1 0 | f = South-East Asia
```

# CLUSTERING



EM -I 100 -N -1 -X 10 -max -1 -ll-cv 1.0E-6 -ll-iter 1.0E-6 -M 1.0E-6 -K 10 -num-slots 1 -S 100

Cluster mode

○ Use training set

○ Supplied test set Set...

○ Percentage split % 66

● Classes to clusters evaluation

(Nom) WHO Region

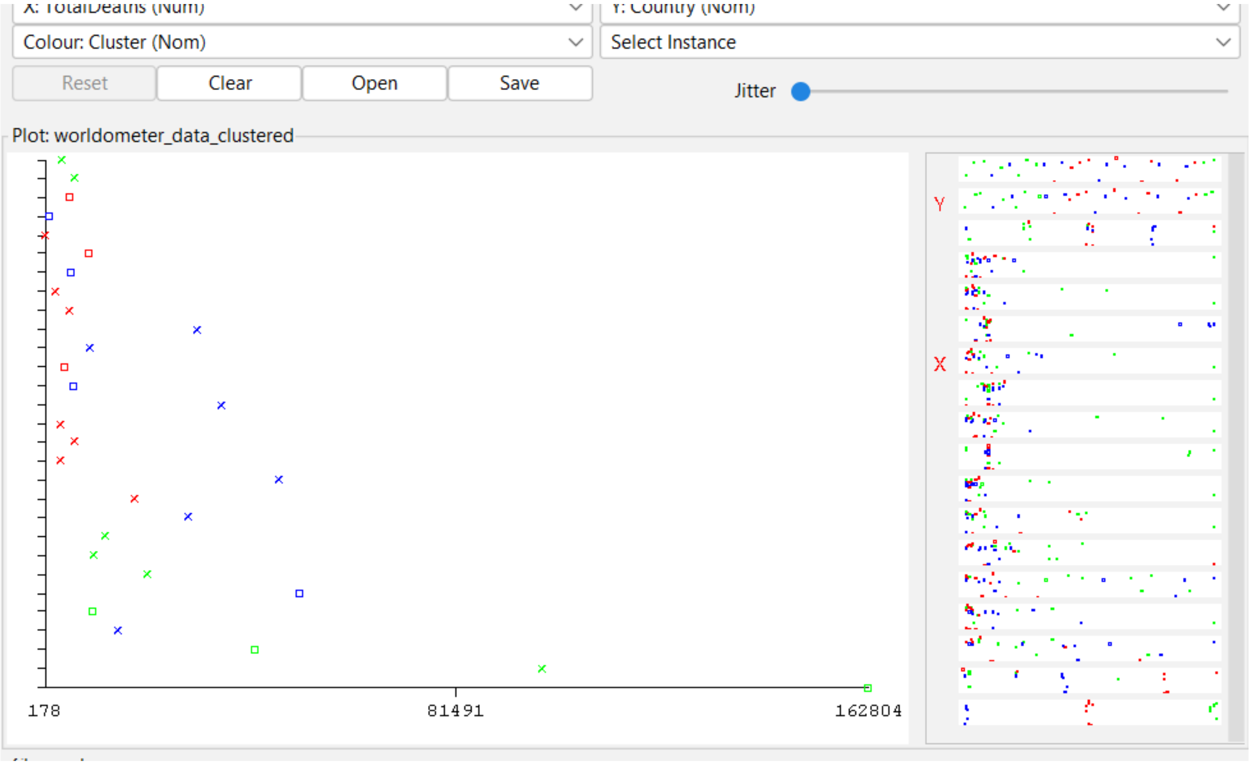
☑ Store clusters for visualization

Result list (right-click for options)

02:58:49 - EM

Clusterer output

```
Time taken to build model (full training data) : 0.15 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       10 ( 34%)
1       10 ( 34%)
2        9 ( 31%)


Log likelihood: -160.77742


Class attribute: Continent
Classes to Clusters:

 0 1 2  <-- assigned to cluster
 1 1 1 | North America
 0 1 6 | South America
 3 7 1 | Asia
 6 0 0 | Europe
 0 1 1 | Africa

Cluster 0 <-- Europe
Cluster 1 <-- Asia
Cluster 2 <-- South America

Incorrectly clustered instances :       10.0     34.4828 %
```



X: TotalDeaths (Num)

Y: Country (Nom)

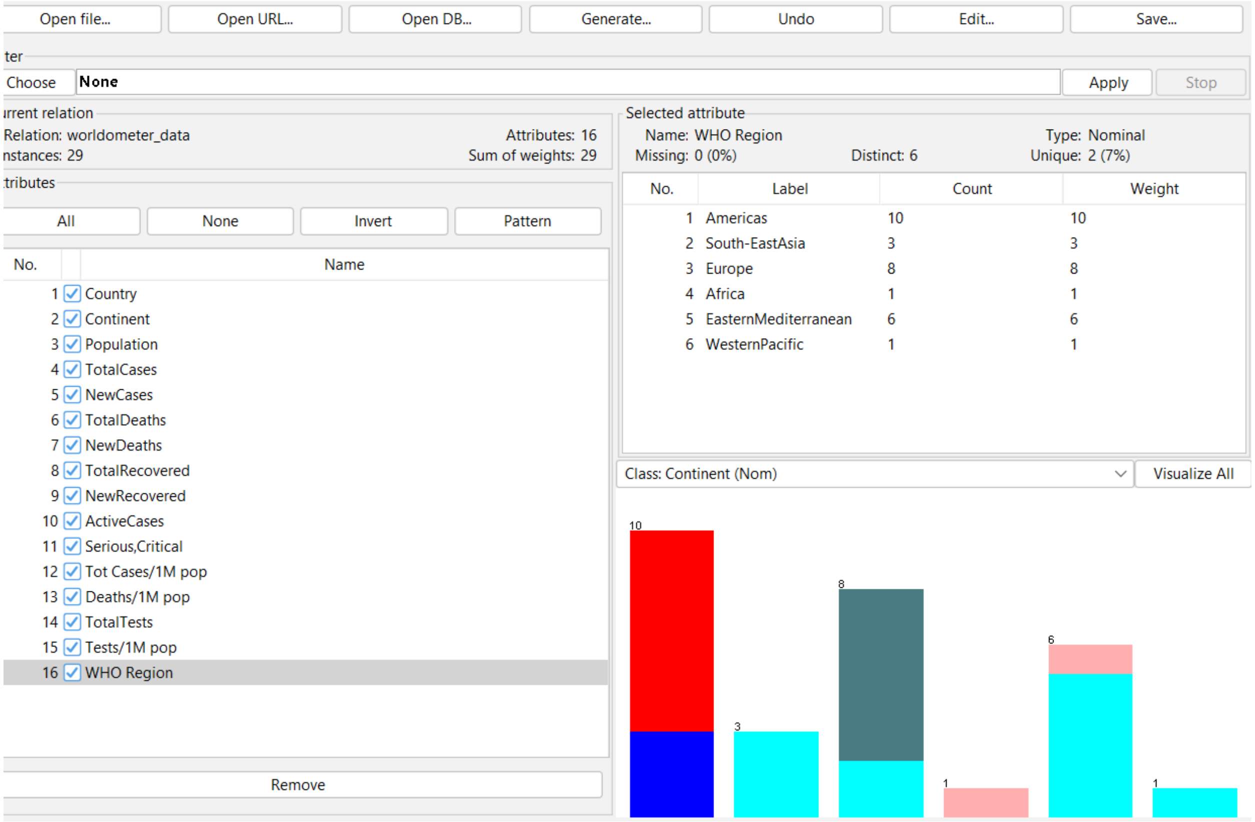
Colour: Cluster (Nom)

Select Instance

Plot: worldometer_data_clustered

# WORLDOMETER_WISE

## LOAD



Choose None

urrent relation

Relation: worldometer_data Attributes: 16

nstances: 29 Sum of weights: 29

tributes

| No. | | Name |
|---|---|---|
| 1 | ☑ | Country |
| 2 | ☑ | Continent |
| 3 | ☑ | Population |
| 4 | ☑ | TotalCases |
| 5 | ☑ | NewCases |
| 6 | ☑ | TotalDeaths |
| 7 | ☑ | NewDeaths |
| 8 | ☑ | TotalRecovered |
| 9 | ☑ | NewRecovered |
| 10 | ☑ | ActiveCases |
| 11 | ☑ | Serious,Critical |
| 12 | ☑ | Tot Cases/1M pop |
| 13 | ☑ | Deaths/1M pop |
| 14 | ☑ | TotalTests |
| 15 | ☑ | Tests/1M pop |
| 16 | ☑ | WHO Region |

Selected attribute

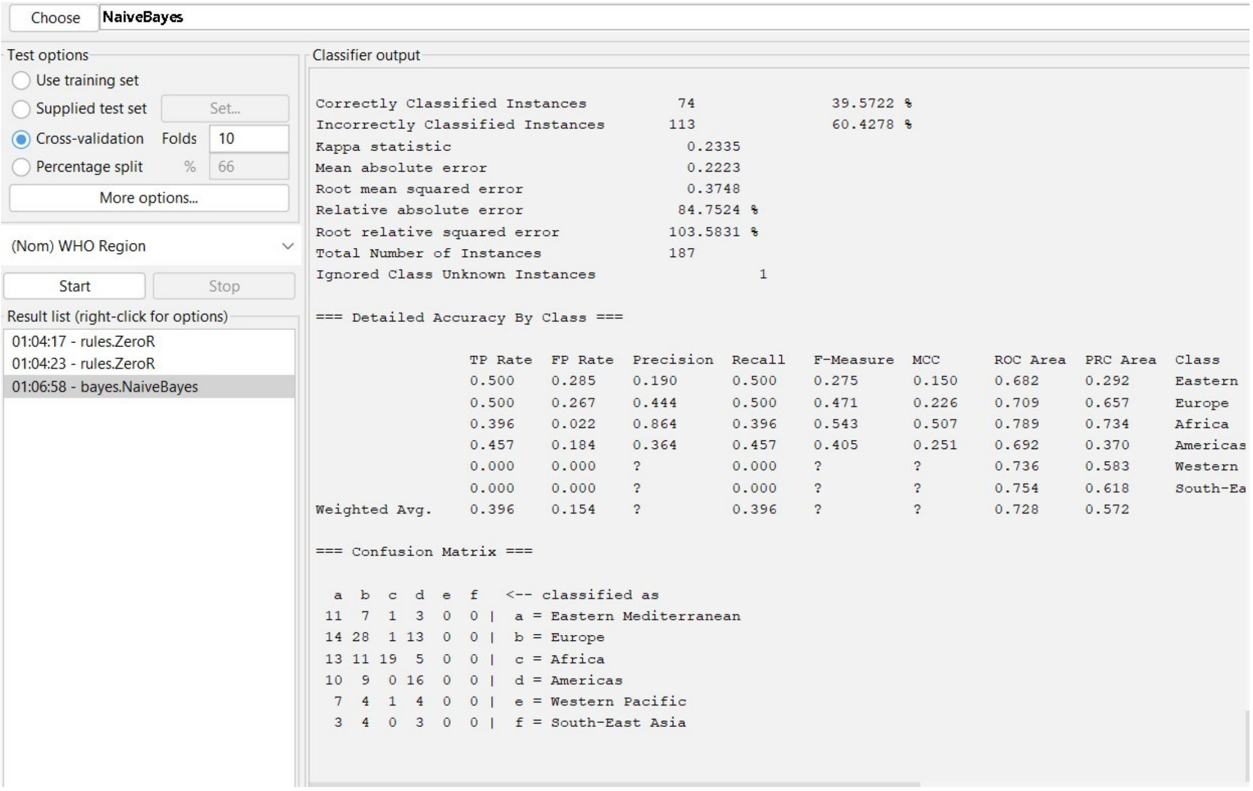
Name: WHO Region Type: Nominal

Missing: 0 (0%) Distinct: 6 Unique: 2 (7%)

| No. | Label | Count | Weight |
|---|---|---|---|
| 1 | Americas | 10 | 10 |
| 2 | South-EastAsia | 3 | 3 |
| 3 | Europe | 8 | 8 |
| 4 | Africa | 1 | 1 |
| 5 | EasternMediterranean | 6 | 6 |
| 6 | WesternPacific | 1 | 1 |

Class: Continent (Nom)

# CLASSIFICATION



NaiveBayes

Test options

○ Use training set

○ Supplied test set

● Cross-validation Folds 10

○ Percentage split % 66

(Nom) WHO Region

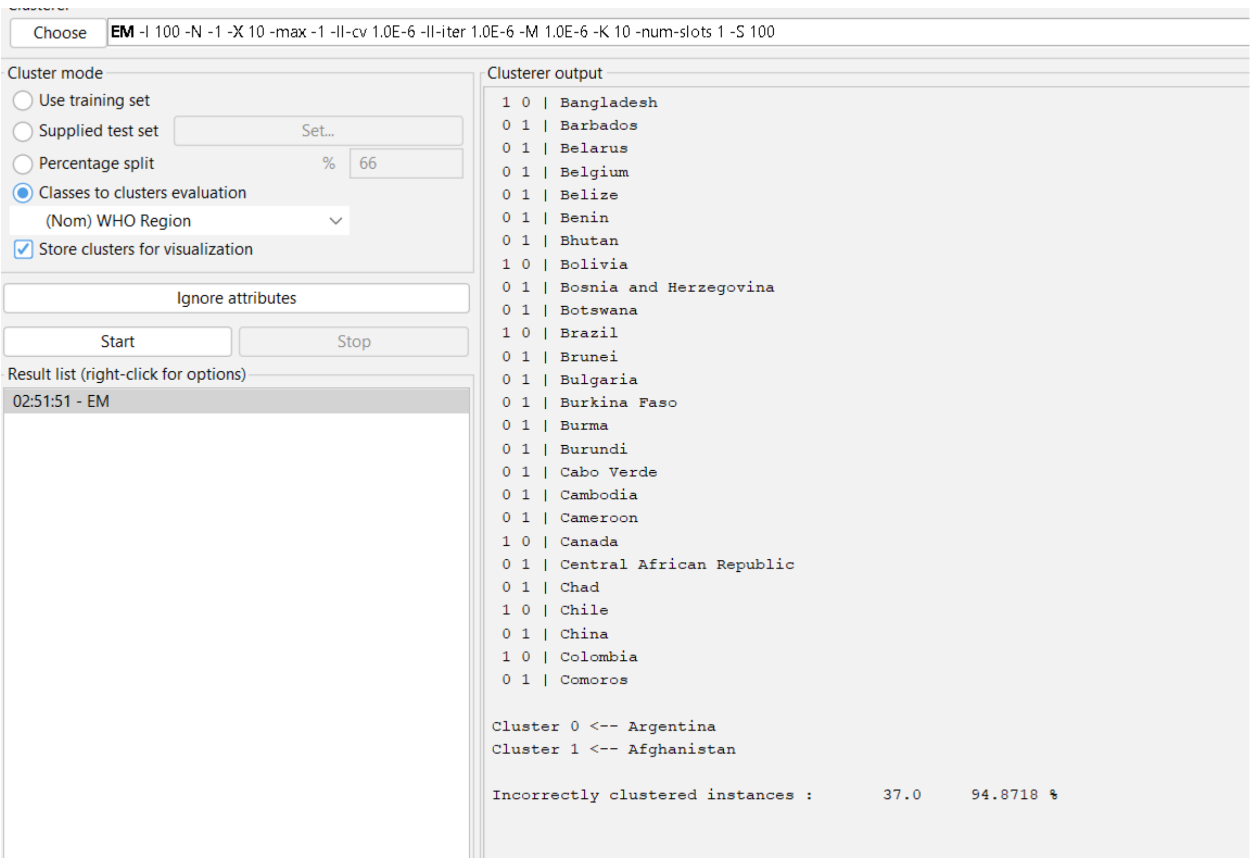
Result list (right-click for options)

01:04:17 - rules.ZeroR
01:04:23 - rules.ZeroR
01:06:58 - bayes.NaiveBayes

Classifier output

```
Correctly Classified Instances          74               39.5722 %
Incorrectly Classified Instances       113               60.4278 %
Kappa statistic                          0.2335
Mean absolute error                      0.2223
Root mean squared error                  0.3748
Relative absolute error                 84.7524 %
Root relative squared error            103.5831 %
Total Number of Instances              187
Ignored Class Unknown Instances                  1

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.500    0.285    0.190      0.500    0.275      0.150    0.682     0.292     Eastern
                 0.500    0.267    0.444      0.500    0.471      0.226    0.709     0.657     Europe
                 0.396    0.022    0.864      0.396    0.543      0.507    0.789     0.734     Africa
                 0.457    0.184    0.364      0.457    0.405      0.251    0.692     0.370     Americas
                 0.000    0.000    ?          0.000    ?          ?        0.736     0.583     Western
                 0.000    0.000    ?          0.000    ?          ?        0.754     0.618     South-Ea
Weighted Avg.    0.396    0.154    ?          0.396    ?          ?        0.728     0.572

=== Confusion Matrix ===

  a  b  c  d  e  f   <-- classified as
 11  7  1  3  0  0 |  a = Eastern Mediterranean
 14 28  1 13  0  0 |  b = Europe
 13 11 19  5  0  0 |  c = Africa
 10  9  0 16  0  0 |  d = Americas
  7  4  1  4  0  0 |  e = Western Pacific
  3  4  0  3  0  0 |  f = South-East Asia
```

# CLUSTERING



```
EM -I 100 -N -1 -X 10 -max -1 -ll-cv 1.0E-6 -ll-iter 1.0E-6 -M 1.0E-6 -K 10 -num-slots 1 -S 100
```

Cluster mode

○ Use training set

○ Supplied test set

○ Percentage split % 66

● Classes to clusters evaluation

(Nom) WHO Region

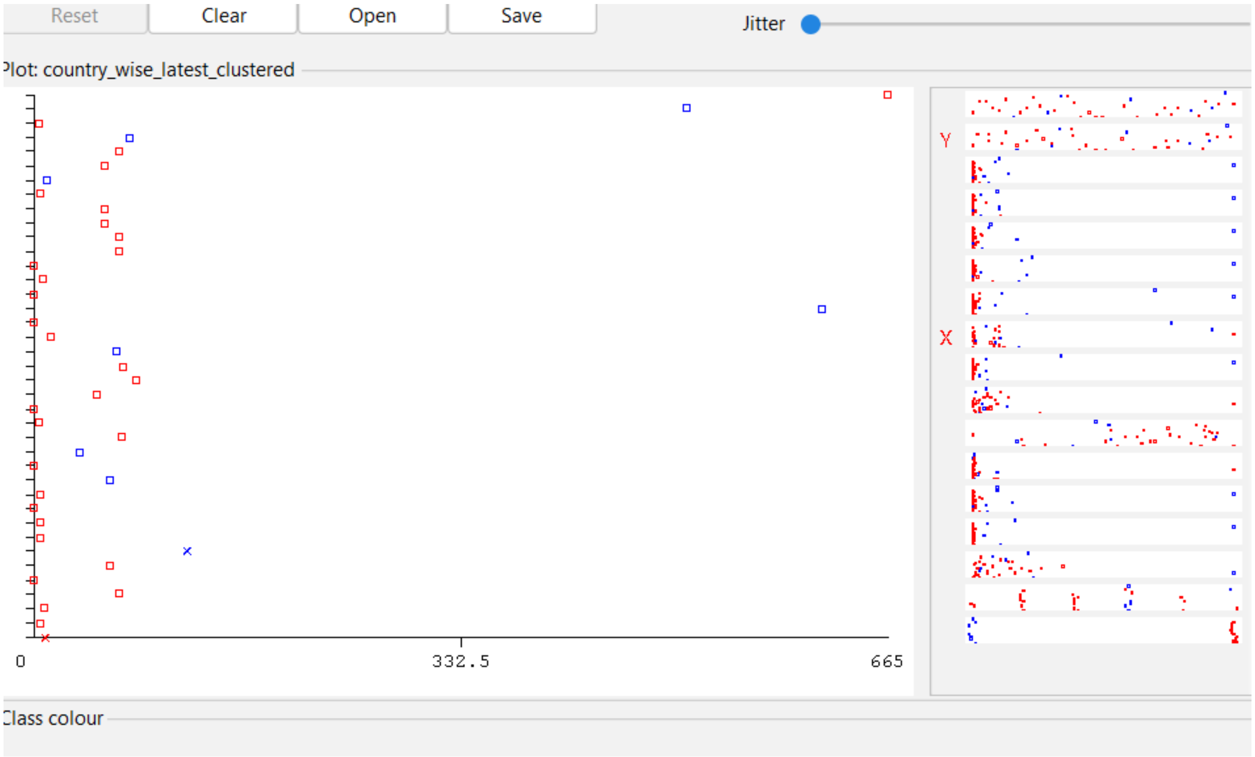
☑ Store clusters for visualization

Result list (right-click for options)

02:51:51 - EM

Clusterer output

```
 1 0 | Bangladesh
 0 1 | Barbados
 0 1 | Belarus
 0 1 | Belgium
 0 1 | Belize
 0 1 | Benin
 0 1 | Bhutan
 1 0 | Bolivia
 0 1 | Bosnia and Herzegovina
 0 1 | Botswana
 1 0 | Brazil
 0 1 | Brunei
 0 1 | Bulgaria
 0 1 | Burkina Faso
 0 1 | Burma
 0 1 | Burundi
 0 1 | Cabo Verde
 0 1 | Cambodia
 0 1 | Cameroon
 1 0 | Canada
 0 1 | Central African Republic
 0 1 | Chad
 1 0 | Chile
 0 1 | China
 1 0 | Colombia
 0 1 | Comoros

Cluster 0 <-- Argentina
Cluster 1 <-- Afghanistan

Incorrectly clustered instances :       37.0     94.8718 %
```



## SUMMARY

We employed an Extract, Transform, Load (ETL) pipeline methodology to comprehensively analyze the progression of new COVID-19 cases and the corresponding recovery rates on a daily basis. This assessment was conducted at both the country-specific and global levels, allowing for a detailed examination of the trends and patterns associated with case recoveries worldwide.