**Data mining using WEKA**

ADITI GOEL 800966155

POWN ARTHI THIMIRI DAYASAGAR 800913286

Rucha Kshirsagar 800956296

SPOORTHY RAVI 800856304

2016

**Table of contents**

|  |  |  |
| --- | --- | --- |
| Sr No | Topic | Page No |
| 1 | Introduction | 2 |
| 2 | Classification  2.1Data  2.2Screens  2.3 Output  2.4 Interpretation | 2  2  4  7  17 |
| 3 | Association Rules | 17 |

1. **Introduction:**

Data mining is one approach that identifies the patterns in data and helps in making decisions by analysis of huge data sets. Weka (**Waikato Environment for Knowledge Analysis** (**Weka**) is a popular suite of machine learning software written in Java, developed at the [University of Waikato](https://en.wikipedia.org/wiki/University_of_Waikato), [New Zealand](https://en.wikipedia.org/wiki/New_Zealand)) .It is a workbench[[1]](https://en.wikipedia.org/wiki/Weka_(machine_learning)#cite_note-1) that contains a collection of visualization tools and algorithms for [data analysis](https://en.wikipedia.org/wiki/Data_analysis) and [predictive modeling](https://en.wikipedia.org/wiki/Predictive_modeling), together with graphical user interfaces for easy access to these functions.

Advantages of Weka include:

* Free availability under the [GNU General Public License](https://en.wikipedia.org/wiki/GNU_General_Public_License).
* Portability, since it is fully implemented in the [Java programming language](https://en.wikipedia.org/wiki/Java_programming_language) and thus runs on almost any modern computing platform.
* A comprehensive collection of data [preprocessing](https://en.wikipedia.org/wiki/Preprocessing) and modeling techniques.
* Ease of use due to its graphical user interfaces.

We have worked on data mining techniques using the Weka software.

1. **Classification:**
   1. **Data**

The raw data for the analysis has been collected from the Website http://fsi.fundfor peace.org/. We have taken the data covering for consecutive years-> **2012,2013,2014,2015** and extended them by adding SIX new features-> Life expectancy at birth total (years), Research and Development, Suicide Rate, Tourism, Global Peace index, Human expenditures

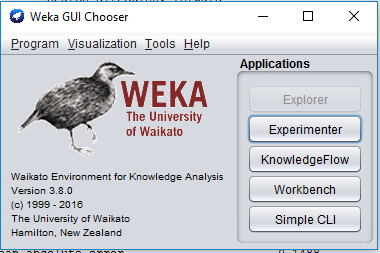
The task is to make sure whether the data has been classified correctly by cross validation.

Below are the files covering data for four consecutive years. These files contains the old attributes as well as the extended attributes we added as part of our analysis.

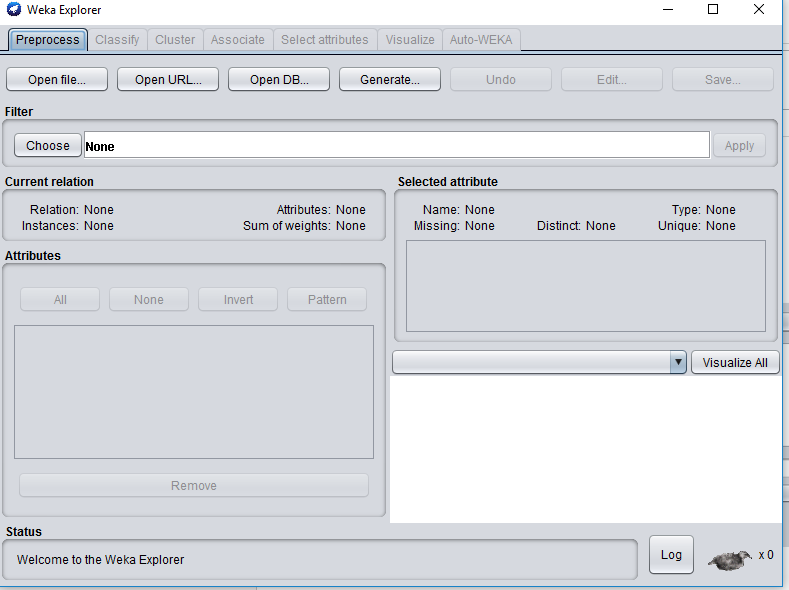
|  |  |  |
| --- | --- | --- |
| **Sr No.** | **Year** | **File** |
| 1. | 2012 |  |
| 2. | 2013 |  |
| 3. | 2014 |  |
| 4. | 2015 |  |

* 1. **Screens**

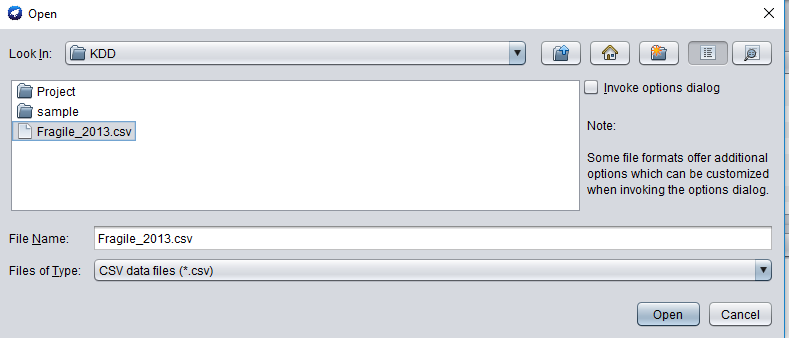
**Step1: Launch Weka Explorer**



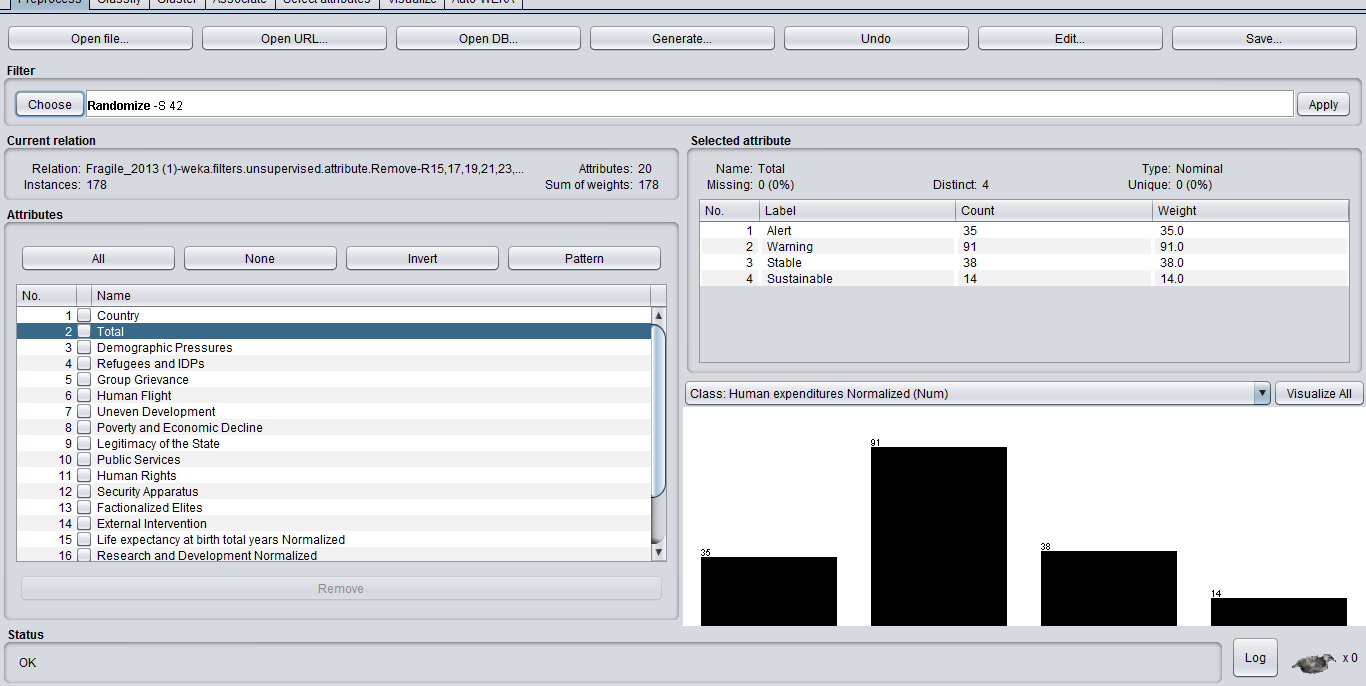
**Step2: Explorer Window**



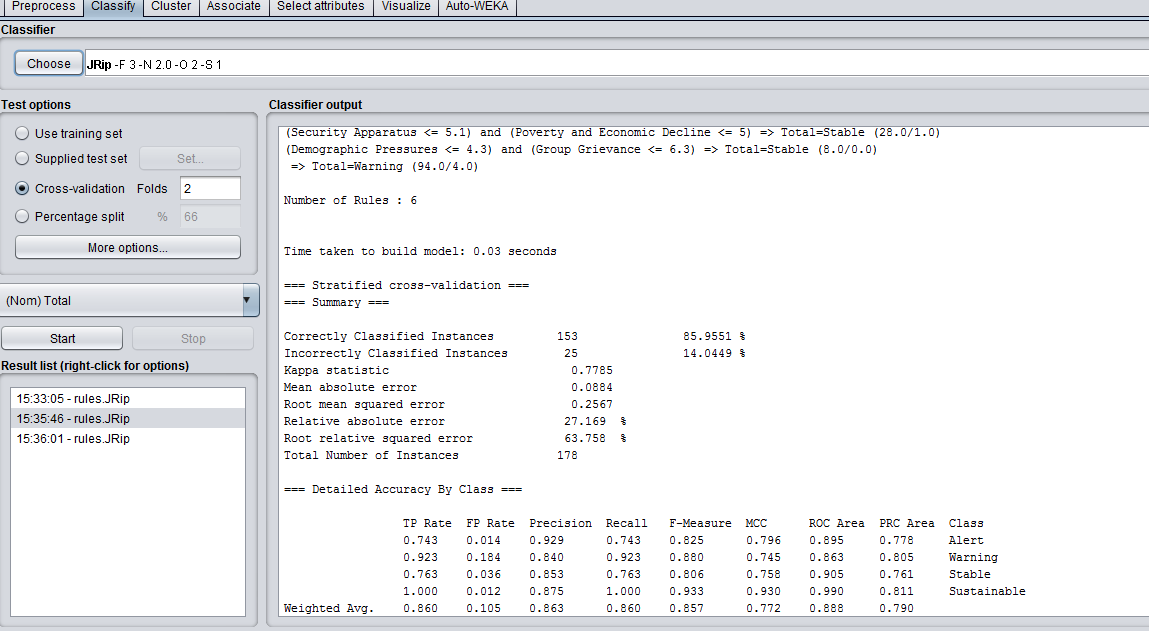
**Step3: Click Open File**



**Step4: Data updated in Weka**



**Step5: Click Cross Validation and choose JRip classifier. Click Start**



**2.3 Output**

**=== Run information ===**

**2012**

Scheme: weka.classifiers.rules.PART -M 2 -C 0.25 -Q 1

Relation: Fragile\_2012

Instances: 178

Attributes:   26

              Fragile States Index 2012

              Demographic Pressures

              Refugees and IDPs

              Group Grievance

              Human Flight

              Uneven Development

              Poverty and Economic Decline

              Legitimacy of the State

              Public Services

              Human Rights

              Security Apparatus

              Factionalized Elites

              External Intervention

              Research and Development expenditure

              RnD Normalize

              Global Peace index

              GPI Normalize

              Life expectancy at birth, total (years)

              Life Epctency Normalize

              Suicide rate

              Suicide Normal

              Tourism

              Tourism Normalize

              Health Expenditures

              Health Normailze

Ouput

Test mode: 20-folds cross-validation

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

Time taken to build model: 0.01 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances 132 74.5763 %

Incorrectly Classified Instances 45 25.4237 %

Kappa statistic 0.5992

Mean absolute error 0.1737

Root mean squared error 0.3217

Relative absolute error 53.9341 %

Root relative squared error 80.2879 %

Total Number of Instances 177

Ignored Class Unknown Instances 1

=== Detailed Accuracy By Class ===

TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class

0.758 0.056 0.758 0.758 0.758 0.702 0.861 0.659 Alert

0.848 0.247 0.788 0.848 0.817 0.605 0.808 0.736 Warning

0.487 0.051 0.731 0.487 0.585 0.511 0.809 0.617 Moderate

0.769 0.055 0.526 0.769 0.625 0.602 0.948 0.442 Sustainable

Weighted Avg. 0.746 0.154 0.750 0.746 0.740 0.602 0.828 0.674

=== Confusion Matrix ===

a b c d <-- classified as

25 8 0 0 | a = Alert

8 78 5 1 | b = Warning

12 19 8 | c = Moderate

1 2 10 | d = Sustainable

**2013**

=== Run information ===

Scheme: weka.classifiers.rules.JRip -F 3 -N 2.0 -O 2 -S 1

Relation: Fragile\_2013 (1)-weka.filters.unsupervised.attribute.Remove-R15,17,19,21,23,25-weka.filters.unsupervised.instance.Randomize-S42

Instances: 178

Attributes: 20

Country

Total

Demographic Pressures

Refugees and IDPs

Group Grievance

Human Flight

Uneven Development

Poverty and Economic Decline

Legitimacy of the State

Public Services

Human Rights

Security Apparatus

Factionalized Elites

External Intervention

Life expectancy at birth total years Normalized

Research and Development Normalized

Suicide Rate Normalized

Tourism Normalized

Global Peace index Normalized

Human expenditures Normalized

Test mode: 5-fold cross-validation

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

JRIP rules:

===========

(Human Rights <= 2) => Total=Sustainable (15.0/2.0)

(Factionalized Elites >= 8.1) and (Poverty and Economic Decline >= 7.3) => Total=Alert (29.0/1.0)

(Security Apparatus >= 8) and (Uneven Development >= 7.2) => Total=Alert (4.0/0.0)

(Security Apparatus <= 5.1) and (Poverty and Economic Decline <= 5) => Total=Stable (28.0/1.0)

(Demographic Pressures <= 4.3) and (Group Grievance <= 6.3) => Total=Stable (8.0/0.0)

=> Total=Warning (94.0/4.0)

Number of Rules : 6

Time taken to build model: 0.03 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances 153 85.9551 %

Incorrectly Classified Instances 25 14.0449 %

Kappa statistic 0.7785

Mean absolute error 0.0884

Root mean squared error 0.2567

Relative absolute error 27.169 %

Root relative squared error 63.758 %

Total Number of Instances 178

=== Detailed Accuracy By Class ===

TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class

0.743 0.014 0.929 0.743 0.825 0.796 0.895 0.778 Alert

0.923 0.184 0.840 0.923 0.880 0.745 0.863 0.805 Warning

0.763 0.036 0.853 0.763 0.806 0.758 0.905 0.761 Stable

1.000 0.012 0.875 1.000 0.933 0.930 0.990 0.811 Sustainable

Weighted Avg. 0.860 0.105 0.863 0.860 0.857 0.772 0.888 0.790

=== Confusion Matrix ===

a b c d <-- classified as

26 9 0 0 | a = Alert

2 84 5 0 | b = Warning

0 7 29 2 | c = Stable

* 1. 0 0 14 | d = Sustainable

**2014**

=== Run information ===

Scheme: weka.classifiers.rules.JRip -F 3 -N 2.0 -O 2 -S 1

Relation: Fragile\_2014edit-weka.filters.unsupervised.attribute.Remove-R3,5,7,9,11,13-weka.filters.unsupervised.instance.Randomize-S42

Instances: 179

Attributes: 20

Country

Total

Life expectancy at birth total years Normalized

Research and Development Normalized

Suicide Rate Normalized

Tourism Normalized

Global Peace Normalized

Health Expenditure Normalized

Demographic Pressures

Refugees and IDPs

Group grievance

Human Flight

Uneven Development

Poverty and Economic Decline

Legitimacy of the State

Public services

Human Rights

Security Apparatus

Factionalized Elites

External Intervention

Test mode: 10-fold cross-validation

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

JRIP rules:

===========

(Life expectancy >= 9.6) and (GlobalPeaceN <= 4) => Total=Sustainable (18.0/5.0)

(R&DN >= 0.8) and (GlobalPeaceN <= 4.9) => Total=Stable (21.0/5.0)

(Life expectancy <= 8.3) => Total=Alert (72.0/17.0)

=> Total=Warning (67.0/16.0)

Number of Rules : 4

Time taken to build model: 0.09 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances 123 69.1011 %

Incorrectly Classified Instances 55 30.8989 %

Kappa statistic 0.5343

Mean absolute error 0.2086

Root mean squared error 0.3525

Relative absolute error 62.028 %

Root relative squared error 86.0629 %

Total Number of Instances 178

Ignored Class Unknown Instances 1

=== Detailed Accuracy By Class ===

TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class

0.803 0.161 0.746 0.803 0.774 0.634 0.818 0.661 Alert

0.681 0.226 0.671 0.681 0.676 0.453 0.723 0.595 Warning

0.407 0.060 0.550 0.407 0.468 0.395 0.789 0.374 Stable

0.769 0.024 0.714 0.769 0.741 0.720 0.882 0.513 Sustainable

Weighted Avg. 0.691 0.162 0.684 0.691 0.685 0.531 0.780 0.580

=== Confusion Matrix ===

a b c d <-- classified as

53 13 0 0 | a = Alert

17 49 6 0 | b = Warning

1 11 11 4 | c = Stable

0 0 3 10 | d = Sustainable

**2015**

=== Run information ===

Scheme: weka.classifiers.trees.J48 -C 0.25 -M 2

Relation: For weka

Instances: 178

Attributes: 26

Country

Total

Demographic Pressures

Refugees and IDPs

Group Grievance

Human Flight

Uneven Development

Poverty and Economic Decline

Legitimacy of the State

Public Services

Human Rights

Security Apparatus

Factionalized Elites

External Intervention

Life expectancy at birth

Life Expectency at Birth - Normalized

Research and Development

R and D - Normalized

Suicide Rates

Suicide Rates - Normalized

Tourism

Tourism - Normalized

Global Peace Index

Global Peace Index - Normalized

Health Expenditures

Health Expenditures - Normalized

Test mode: 2-fold cross-validation

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

J48 pruned tree

------------------

Security Apparatus <= 4.7

| Legitimacy of the State <= 1.8

| | Global Peace Index <= 1.5: Sustainable (16.0/1.0)

| | Global Peace Index > 1.5: Stable (2.0)

| Legitimacy of the State > 1.8

| | Demographic Pressures <= 6.2

| | | External Intervention <= 6: Stable (31.0)

| | | External Intervention > 6: Warning (3.0/1.0)

| | Demographic Pressures > 6.2: Warning (2.0)

Security Apparatus > 4.7

| Refugees and IDPs <= 7.6

| | Factionalized Elites <= 8.2

| | | Factionalized Elites <= 5.6

| | | | Demographic Pressures <= 5.1: Stable (2.0)

| | | | Demographic Pressures > 5.1: Warning (13.0/1.0)

| | | Factionalized Elites > 5.6: Warning (59.0)

| | Factionalized Elites > 8.2

| | | Poverty and Economic Decline <= 7.5

| | | | Demographic Pressures <= 7.8: Warning (8.0)

| | | | Demographic Pressures > 7.8: Alert (2.0)

| | | Poverty and Economic Decline > 7.5: Alert (5.0)

| Refugees and IDPs > 7.6

| | Life expectancy at birth <= 66.2: Alert (28.0)

| | Life expectancy at birth > 66.2

| | | Group Grievance <= 8.9: Warning (4.0)

| | | Group Grievance > 8.9: Alert (3.0)

Number of Leaves : 14

Size of the tree : 27

Time taken to build model: 0.01 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances 152 85.3933 %

Incorrectly Classified Instances 26 14.6067 %

Kappa statistic 0.7763

Mean absolute error 0.0867

Root mean squared error 0.2657

Relative absolute error 26.0124 %

Root relative squared error 65.2503 %

Total Number of Instances 178

=== Detailed Accuracy By Class ===

TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class

0.842 0.036 0.865 0.842 0.853 0.814 0.927 0.769 Alert

0.908 0.143 0.859 0.908 0.883 0.765 0.886 0.834 Warning

0.789 0.050 0.811 0.789 0.800 0.747 0.877 0.709 Stable

0.733 0.006 0.917 0.733 0.815 0.806 0.950 0.768 Sustainable

Weighted Avg. 0.854 0.089 0.855 0.854 0.853 0.775 0.898 0.788

=== Confusion Matrix ===

a b c d <-- classified as

32 6 0 0 | a = Alert

5 79 3 0 | b = Warning

0 7 30 1 | c = Stable

0 0 4 11 | d = Sustainable

**2.4 Interpretation of the Results**

* **2012-** There are 74.5763 % correctly classified instances and 25.4237 % incorrectly classified instances.
* **2013-** There are 85.9551 % correctly classified instances and 14.0449 % incorrectly classified instances.
* **2014-** There are 69.1011 % correctly classified instances and 30.8989 % incorrectly classified instances.
* **2015-** There are 85.3933 % correctly classified instances and 14.6067 % incorrectly classified instances.

**3.Association rules**

The association rules were generated based on Apriori Algorithm.

**2012**

=== Run information ===

Scheme:       weka.associations.Apriori -N 10 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1 -S -1.0 -c -1

Relation:     Fragile\_2012-weka.filters.unsupervised.attribute.NumericToNominal-Rfirst-last

Instances:    178

Attributes:   26

              Fragile States Index 2012

              Demographic Pressures

              Refugees and IDPs

              Group Grievance

              Human Flight

              Uneven Development

              Poverty and Economic Decline

              Legitimacy of the State

              Public Services

              Human Rights

              Security Apparatus

              Factionalized Elites

              External Intervention

              Research and Development expenditure

              RnD Normalize

              Global Peace index

              GPI Normalize

              Life expectancy at birth, total (years)

              Life Epctency Normalize

              Suicide rate

              Suicide Normal

              Tourism

              Tourism Normalize

              Health Expenditures

              Health Normailze

              Ouput

=== Associator model (full training set) ===

Apriori

=======

Minimum support: 0.15 (27 instances)

Minimum metric <confidence>: 0.9

Number of cycles performed: 17

Generated sets of large itemsets:

Size of set of large itemsets L(1): 7

Size of set of large itemsets L(2): 8

Size of set of large itemsets L(3): 3

Best rules found:

 1. RnD Normalize=0 95 ==> Research and Development expenditure=0 95    <conf:(1)> lift:(1.8) lev:(0.24) [42] conv:(42.16)

 2. RnD Normalize=0 Ouput=Warning 58 ==> Research and Development expenditure=0 58    <conf:(1)> lift:(1.8) lev:(0.14) [25] conv:(25.74)

 3. RnD Normalize=0 Tourism Normalize=0 32 ==> Research and Development expenditure=0 32    <conf:(1)> lift:(1.8) lev:(0.08) [14] conv:(14.2)

 4. RnD Normalize=0 Ouput=Alert 30 ==> Research and Development expenditure=0 30    <conf:(1)> lift:(1.8) lev:(0.07) [13] conv:(13.31)

 5. Research and Development expenditure=0 Ouput=Alert 30 ==> RnD Normalize=0 30    <conf:(1)> lift:(1.87) lev:(0.08) [13] conv:(13.99)

 6. Research and Development expenditure=0 99 ==> RnD Normalize=0 95    <conf:(0.96)> lift:(1.8) lev:(0.24) [42] conv:(9.23)

 7. Research and Development expenditure=0 Tourism Normalize=0 34 ==> RnD Normalize=0 32    <conf:(0.94)> lift:(1.76) lev:(0.08) [13] conv:(5.28)

 8. Research and Development expenditure=0 Ouput=Warning 62 ==> RnD Normalize=0 58    <conf:(0.94)> lift:(1.75) lev:(0.14) [24] conv:(5.78)

 9. Ouput=Alert 33 ==> Research and Development expenditure=0 30    <conf:(0.91)> lift:(1.63) lev:(0.07) [11] conv:(3.66)

10. Ouput=Alert 33 ==> RnD Normalize=0 30    <conf:(0.91)> lift:(1.7) lev:(0.07) [12] conv:(3.85)

**2013**

**===** Run information **===**

Scheme: weka.associations.Apriori -N 10 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1 -S -1.0 -c -1

Relation: Fragile\_2013 (1)-weka.filters.unsupervised.attribute.Remove-R15,17,19,21,23,25-weka.filters.unsupervised.attribute.NumericToNominal-Rfirst-last

Instances: 178

Attributes: 20

Country

Total

Demographic Pressures

Refugees and IDPs

Group Grievance

Human Flight

Uneven Development

Poverty and Economic Decline

Legitimacy of the State

Public Services

Human Rights

Security Apparatus

Factionalized Elites

External Intervention

Life expectancy at birth total years Normalized

Research and Development Normalized

Suicide Rate Normalized

Tourism Normalized

Global Peace index Normalized

Human expenditures Normalized

=== Associator model (full training set) ===

Apriori

=======

Minimum support: 0.1 (18 instances)

Minimum metric <confidence>: 0.9

Number of cycles performed: 18

Generated sets of large itemsets:

Size of set of large itemsets L(1): 8

Size of set of large itemsets L(2): 7

Size of set of large itemsets L(3): 2

Best rules found:

1. Total=Alert Tourism Normalized=0 20 ==> Research and Development Normalized=0 20 <conf:(1)> lift:(1.71) lev:(0.05) [8] conv:(8.31)

2. Total=Alert 35 ==> Research and Development Normalized=0 34 <conf:(0.97)> lift:(1.66) lev:(0.08) [13] conv:(7.28)

**2014**

=== Run information ===

Scheme: weka.associations.Apriori -N 10 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1 -S -1.0 -c -1

Relation: Fragile\_2014edit-weka.filters.unsupervised.attribute.Remove-R3,5,7,9,11,13-weka.filters.unsupervised.instance.Randomize-S42-weka.filters.unsupervised.attribute.NumericToNominal-Rfirst-last

Instances: 179

Attributes: 20

Country

Total

Life expectancy at birth total years Normalized

Research and Development Normalized

Suicide Rate Normalized

Tourism Normalized

Global Peace Normalized

Health Expenditure Normalized

Demographic Pressures

Refugees and IDPs

Group grievance

Human Flight

Uneven Development

Poverty and Economic Decline

Legitimacy of the State

Public services

Human Rights

Security Apparatus

Factionalized Elites

External Intervention

=== Associator model (full training set) ===

Apriori

=======

Minimum support: 0.1 (18 instances)

Minimum metric <confidence>: 0.9

Number of cycles performed: 18

Generated sets of large itemsets:

Size of set of large itemsets L(1): 8

Size of set of large itemsets L(2): 6

Size of set of large itemsets L(3): 1

Best rules found:

1. Total=Alert TourismN=0 39 ==> R&DN=0 37 <conf:(0.95)> lift:(1.63) lev:(0.08) [14] conv:(5.45)

\

**2015**

=== Run information ===

Scheme: weka.associations.Apriori -N 10 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1 -S -1.0 -c -1

Relation: For weka-weka.filters.unsupervised.attribute.NumericToNominal-Rfirst-last

Instances: 178

Attributes: 26

Country

Total

Demographic Pressures

Refugees and IDPs

Group Grievance

Human Flight

Uneven Development

Poverty and Economic Decline

Legitimacy of the State

Public Services

Human Rights

Security Apparatus

Factionalized Elites

External Intervention

Life expectancy at birth

Life Expectency at Birth - Normalized

Research and Development

R and D - Normalized

Suicide Rates

Suicide Rates - Normalized

Tourism

Tourism - Normalized

Global Peace Index

Global Peace Index - Normalized

Health Expenditures

Health Expenditures - Normalized

=== Associator model (full training set) ===

Apriori

=======

Minimum support: 0.2 (36 instances)

Minimum metric <confidence>: 0.9

Number of cycles performed: 16

Generated sets of large itemsets:

Size of set of large itemsets L(1): 6

Size of set of large itemsets L(2): 7

Size of set of large itemsets L(3): 3

Best rules found:

1. R and D - Normalized=0 113 ==> Research and Development=0 113 <conf:(1)> lift:(1.56) lev:(0.23) [40] conv:(40.63)

2. Total=Warning R and D - Normalized=0 63 ==> Research and Development=0 63 <conf:(1)> lift:(1.56) lev:(0.13) [22] conv:(22.65)

3. R and D - Normalized=0 Tourism - Normalized=0 45 ==> Research and Development=0 45 <conf:(1)> lift:(1.56) lev:(0.09) [16] conv:(16.18)

4. Research and Development=0 Tourism - Normalized=0 45 ==> R and D - Normalized=0 45 <conf:(1)> lift:(1.58) lev:(0.09) [16] conv:(16.43)

5. Total=Alert R and D - Normalized=0 37 ==> Research and Development=0 37 <conf:(1)> lift:(1.56) lev:(0.07) [13] conv:(13.3)

6. Total=Alert Research and Development=0 37 ==> R and D - Normalized=0 37 <conf:(1)> lift:(1.58) lev:(0.08) [13] conv:(13.51)

7. Research and Development=0 114 ==> R and D - Normalized=0 113 <conf:(0.99)> lift:(1.56) lev:(0.23) [40] conv:(20.81)

8. Total=Warning Research and Development=0 64 ==> R and D - Normalized=0 63 <conf:(0.98)> lift:(1.55) lev:(0.13) [22] conv:(11.69)

9. Total=Alert 38 ==> Research and Development=0 37 <conf:(0.97)> lift:(1.52) lev:(0.07) [12] conv:(6.83)

10. Total=Alert 38 ==> R and D - Normalized=0 37 <conf:(0.97)> lift:(1.53) lev:(0.07) [12] conv:(6.94)