Analysis on Fragile States

Data- Department of Computer Science

University of North Carolina at Charlotte,NC.

Instructor Dr.Zbigniew W. Ras

By:

Md Hussain Musthaq - #801155825

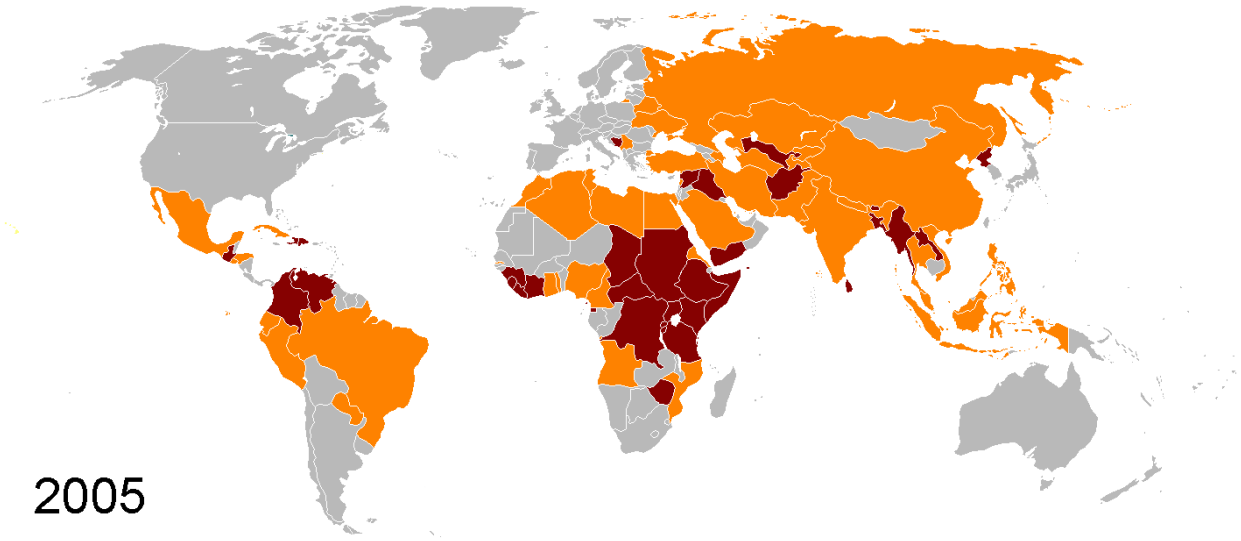
Gokul Mani - #801133944

Vaishali Krishnamurthy - #801077752

I. **Introduction**:

The thought process of this venture is to perform Action-Rule digging for given Fragile State Index Datasets with new stretched out highlights added to the informational index and further dissect how the activity rules change throughout the years. Utilizing activity rules, we survey how a Country can move from the province of Alert to Stable State. This report additionally talks about the significance of each quality in informational index and how they are answerable for choosing the state for the nation. Discretization and Classification of information is additionally accomplished for additional investigation.

II. **Fragile State Index**



The Fragile States Index (FSI) is a yearly positioning of 178 nations dependent on the various weights they face that sway their degrees of delicacy. The Index depends on The Fund for Peace's exclusive Conflict Assessment System Tool (CAST) expository methodology. In view of thorough sociology approach, three essential floods of information quantitative, subjective, and master approval - are triangulated and exposed to basic audit to acquire last scores for the FSI. A large number of archives are investigated each year, and by applying profoundly particular hunt parameters, scores are allotted for each nation dependent on twelve key political, social and monetary pointers and more than 100 sub-markers that are the aftereffect of long stretches of master sociology inquire about.

The FSI is the entirety of scores of twelve particular highlights (to be specific, Demographic weights, Refugees and IDPs, Group Grievance, Human Flight, Uneven Development,

Neediness and Economic Decline, State Legitimacy, Public Services, Human Rights, Security Apparatus, Factionalized Elites, External Intervention) demonstrative of state's soundness. These highlights are extensively classified into social, monetary and political gatherings. Each element has its worth characterized inside the range of 0 and 10, with a higher worth showing a higher

level of delicacy, in this manner making a scale going 0−120. Every one of these highlights is clarified beneath:

III. Data Set:

a) Dataset Features:

- **Social Indicators**

1. Demographic Pressures (DP): Characterized by the overall proportion of high populace thickness and other life sustaining assets like nourishment, water and so forth.

2. Refugees and IDP: Characterized by the movement of all-inclusive community that can influence security and different assets.

3. Group Grievance: Characterized by the current viciousness between networks that can compromise the nation's security.

4. Human Flight: Characterized by relocation per capita particularly migration of instructed individuals who moves out of nations for better chance.

- **Economic Indicators**

5. Poverty and Economic Decline: Characterized by the monetary decay of the general public all in all that consequently influences the nation's capacity to give better result.

6. Uneven Development: Characterized by the imbalance in various classifications like work, instruction, monetary that sway contrarily on the implicit understanding of a nation.

- **Political Indicators**

7. State Legitimacy: Characterized by the defilement in the more significant level of a state's locale.

8. Public Services: Characterized by the absence of significant administrations like medicinal services, instruction, tidiness, transportation that influences the exhibition of a state.

9. Human Rights: Characterized by the frail assurance of fundamental rights demonstrates the disappointment of a state to play out its essential duty.

10. Security Apparatus: A rise of first class or praetorian monitors that work without risk of punishment challenges the security mechanical assembly's imposing business model on the utilization of power, 4 debilitating the implicit understanding. Estimations incorporate inward clash, riots/fights, military overthrows, rebel action.

11. Factionalized Elites: Measured by power input from elite groups and corrupted elections.

12. External Intervention: Estimated by outside help, nearness of peacekeepers or UN missions, remote military mediation, authorizes, and FICO assessments.

b) Extended Data

13. Literacy Rate Adult: This feature counts the literacy rate among the adults in the given country.

14.Export and Imports: This feature allows to monitor the exports and imports done by the country with outer states.

15.Foreign Direct Investment Net inflows: Here we are defining net investment on a country. The range is defined between 0-100

16. Life Expectancy at Birth: The mortality rate of the infants is taken in account here. i.e. Probability of infant that dies within one year of birth. Which is expressed as thousand live births.

17.Unemployment Total: This feature tracks the unemployment factor in the country.

18. Life Expectancy Index: Number of years a newborn infant could expect to live if prevailing patterns of age-specific mortality rates at the time of birth.

c) Data Extraction

The values for these extended features have been extracted from various websites and added to the original FSI dataset. The data from the year 2010 is compiled and stored as an Excel sheet for further processing & analysis.

Values for the extended features were gathered from various websites which are listed below:

- http://hdr.undp.org/en/
- https://data.worldbank.org/

IV. Data Cleaning and Preprocessing

The data obtained from the FSI website was not clean and hence some pre-processing was done before using it for classification. Following were the steps taken for Data Cleaning.

1. Special Characters like !,% etc. were removed so that the sheets can be parsed by WEKA tool.
2. Missing numeric values were filled by their mean values.
3. Outliers were removed and replaced by mean values.
4. Rows having lot of empty values were removed from the analysis.
5. Converting each numeric column value in the range of 0 to 10.
6. Numeric values of Decision Variable Total were replaced by Nominal Values based on the following table values.

DISCRETIZATION:

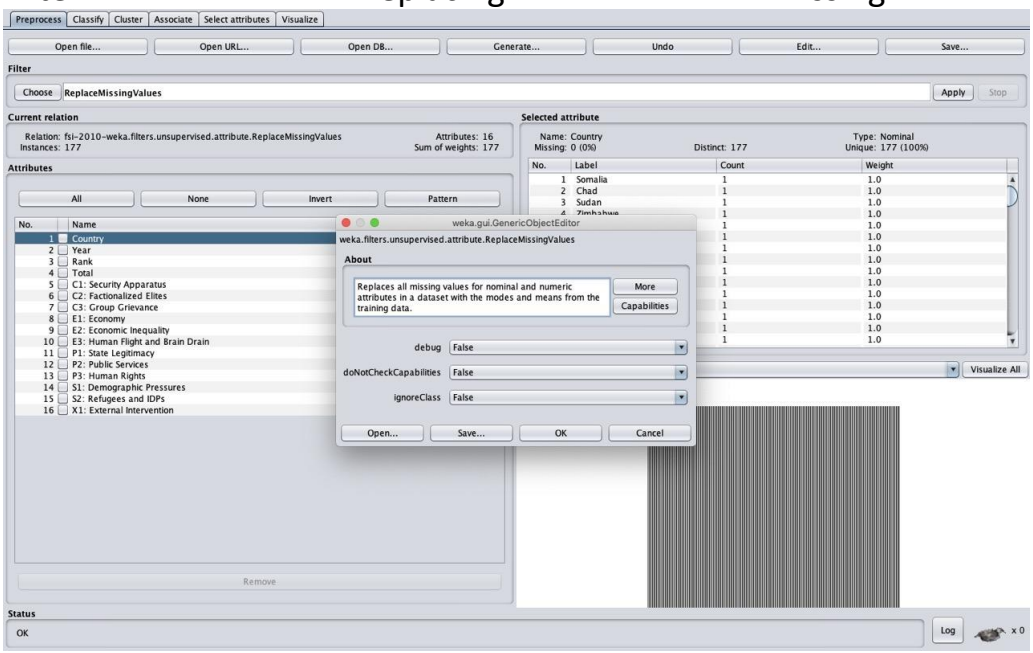
Discretization refers to the process of converting or partitioning continuous attributes, features or variables to discretized or nominal attributes/features/variables. The decision attribute TOTAL has various values throughout the dataset, we have used discretization to replace numeric values of TOTAL by the following concepts.

1. Alert (union of Very High Alert, High Alert, Alert)
2. Warning (union of High Warning, Elevated Warning, Warning).
3. Stable (union of Stable, More Stable, Very Stable).
4. Sustainable (union of Sustainable, Very Sustainable).

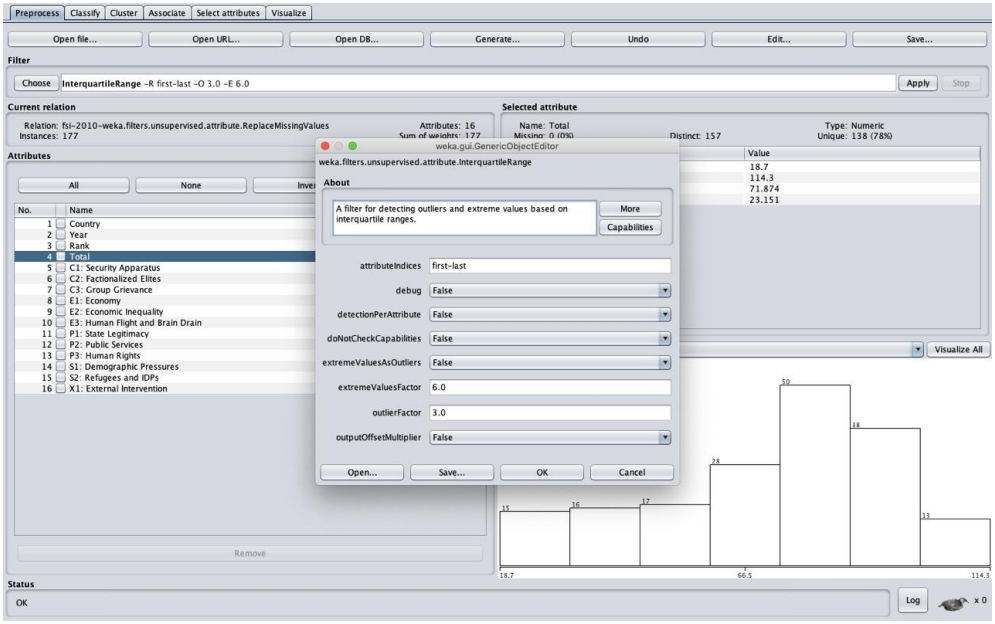
Year-2010

Missing Values:

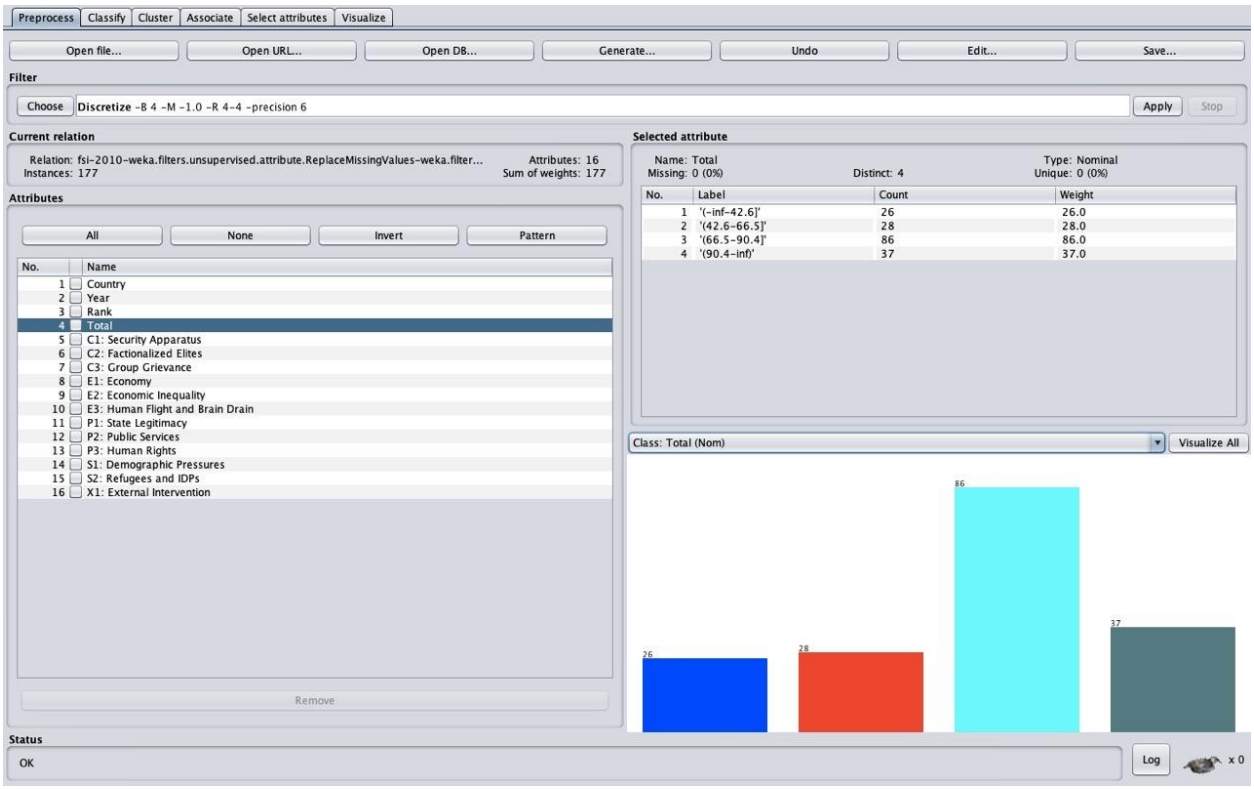
Filter- Replacing Missing Values

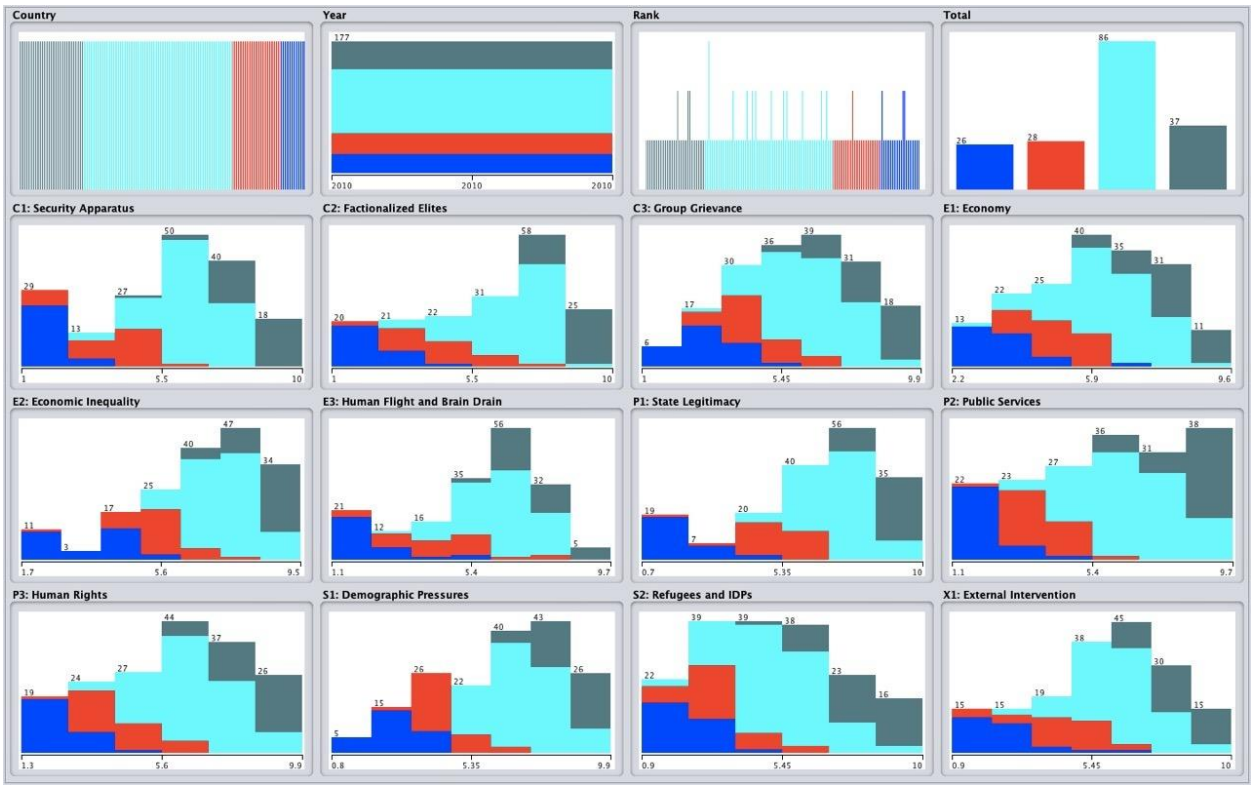


OUTLIERREMOVAL:
Filter-InterQuartileRange

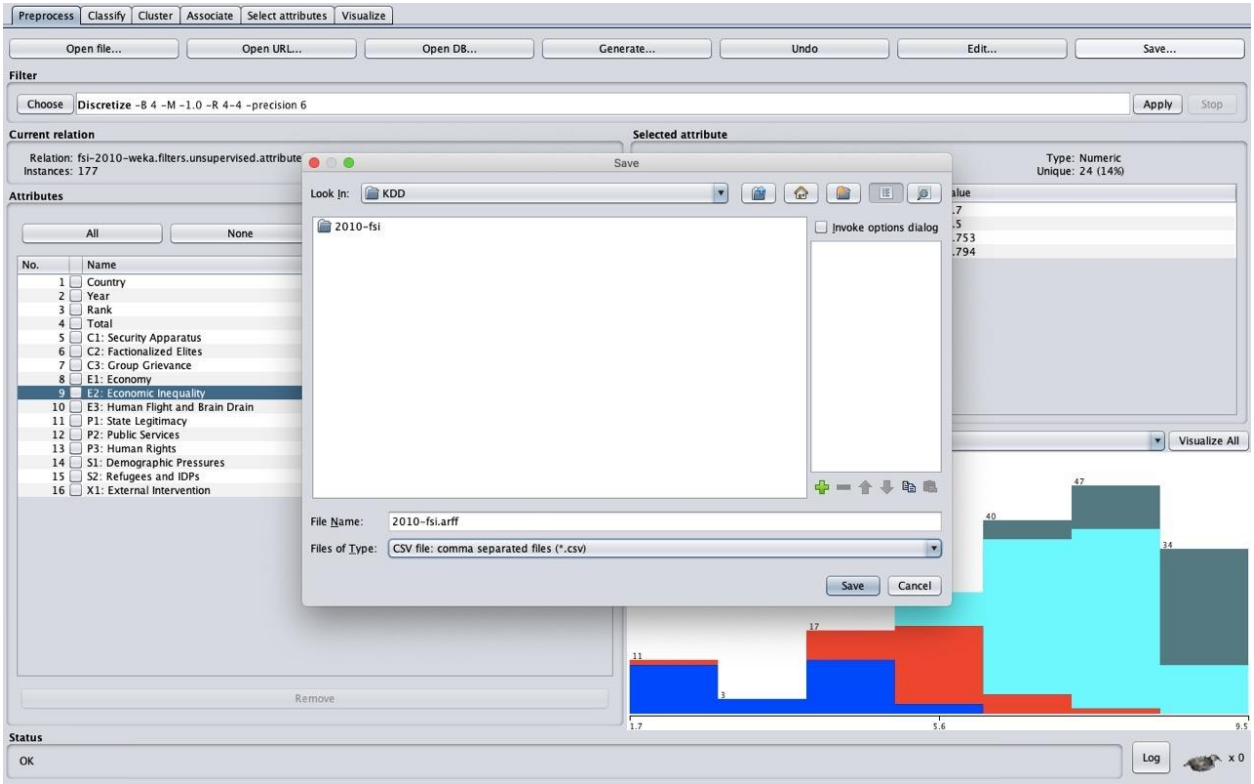


DISCRETIZATION
FILTER -DISCRETIZE





Final Save to CSV



V Classification

For Classification, we have utilized the accompanying classifiers. Every classifier is run for the dataset year 2010. Also, we have calculated the accuracy for classifier for the datasets before adding the extra-features and after adding the extra-features to compare the efficiency of both types of datasets. Following are the Algorithms used for Classification:

1) Naïve Bayes: Class for a Naive Bayes classifier using estimator classes. Numeric estimator precision values are chosen based on analysis of the training data. For this reason, the classifier is not an Updateable Classifier (which in typical usage are initialized with zero training instances) -- if you need the Updateable Classifier functionality, use the NaiveBayes Updateable classifier. The NaiveBayes Updateable classifier will use a default precision of 0.1 for numeric attributes when buildClassifier is called with zero training instances.

2) J48: The C4.5 algorithm for building decision trees is implemented in Weka as a classifier called J48. Classifiers, like filters, are organized in a hierarchy: J48 has the full name weka.classifiers.trees.J48. 10

3) JRip: 1.1. Grow phase: This class implements a propositional rule learner, Repeated Incremental Pruning to Produce Error Reduction (RIPPER), which was proposed by William W. Cohen as an optimized version of IREP.

The algorithm is briefly described as follows:

Initialize RS = {}, and for each class from the less prevalent one to the more frequent one, DO:

a). Building stage:

Repeat 1.1 and 1.2 until the description length (DL) of the ruleset and examples is 64 bits greater than the smallest DL met so far, or there are no positive examples, or the error rate >= 50%.

Grow one rule by greedily adding antecedents (or conditions) to the rule until the rule is perfect (i.e. 100% accurate). The procedure tries every possible value of each attribute and selects the condition with highest information gain: p(log(p/t)-log(P/T)).

1.2. Prune phase:

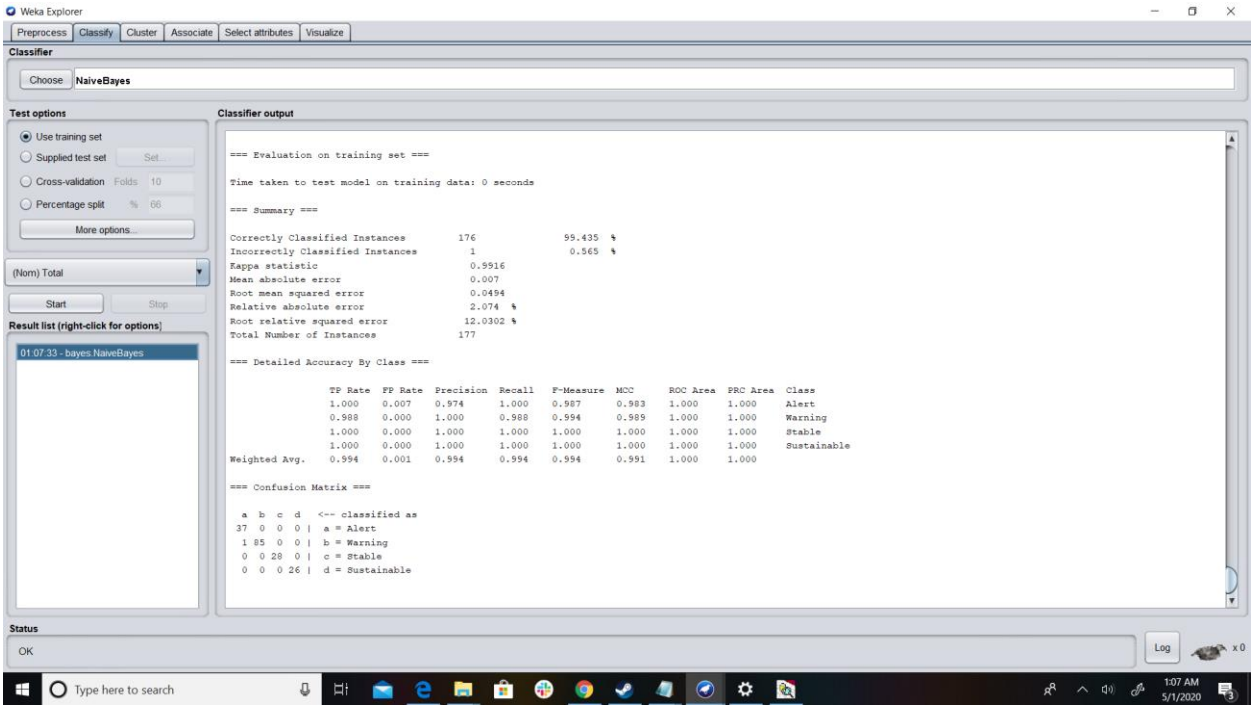
Incrementally prune each rule and allow the pruning of any final sequences of the antecedents;The pruning metric is (p-n)/(p+n) -- but it's actually 2p/(p+n) -1, so in this implementation we simply use p/(p+n) (actually (p+1)/(p+n+2), thus if p+n is 0, it's 0.5).

b). Optimization stage:

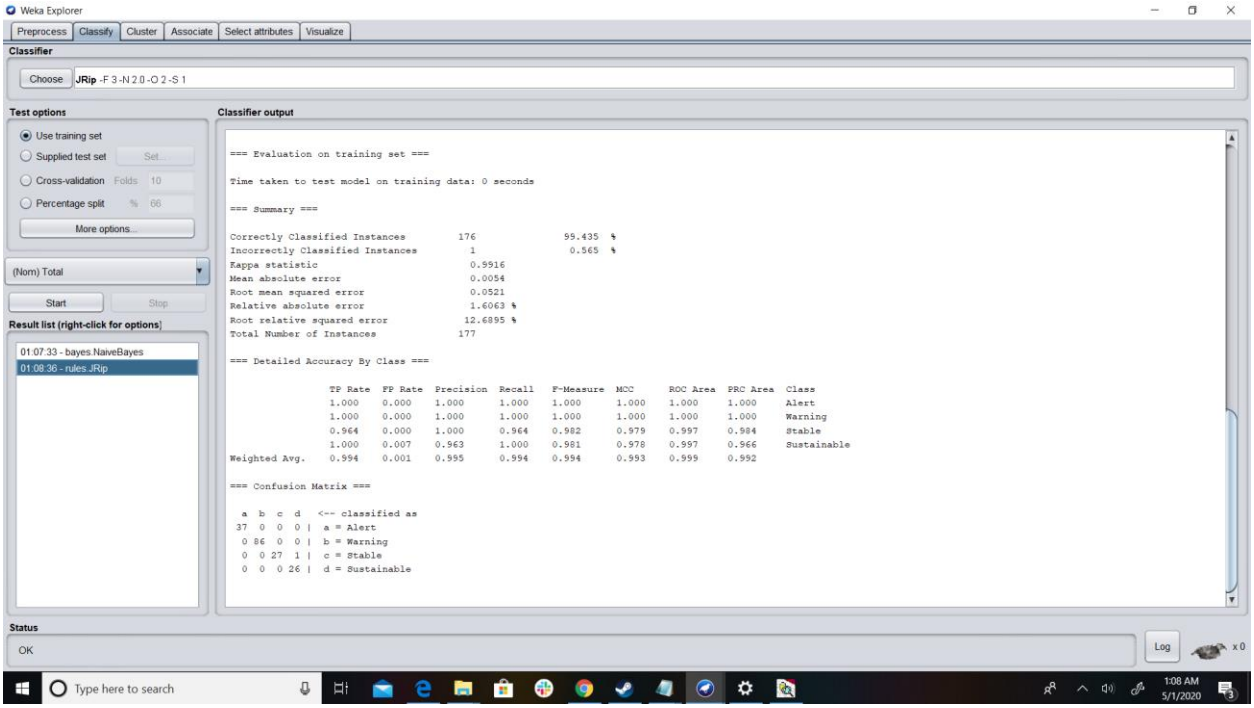
After generating the initial ruleset {Ri}, generate and prune two variants of each rule Ri from randomized data using procedure 1.1 and 1.2. But one variant is generated from an empty rule while the other is generated by greedily adding antecedents to the original rule. Moreover, the pruning metric used here is (TP+TN)/(P+N).Then the smallest possible DL for 11 each variant and the original rule is computed. The variant with the minimal DL is selected as the final representative of Ri in the ruleset.After all the rules in {Ri} have been examined and if there are still residual positives, more rules are generated based on the residual positives using Building Stage again.

c). Delete the rules from the ruleset that would increase the DL of the whole ruleset if it were in it. and add resultant ruleset to RS.

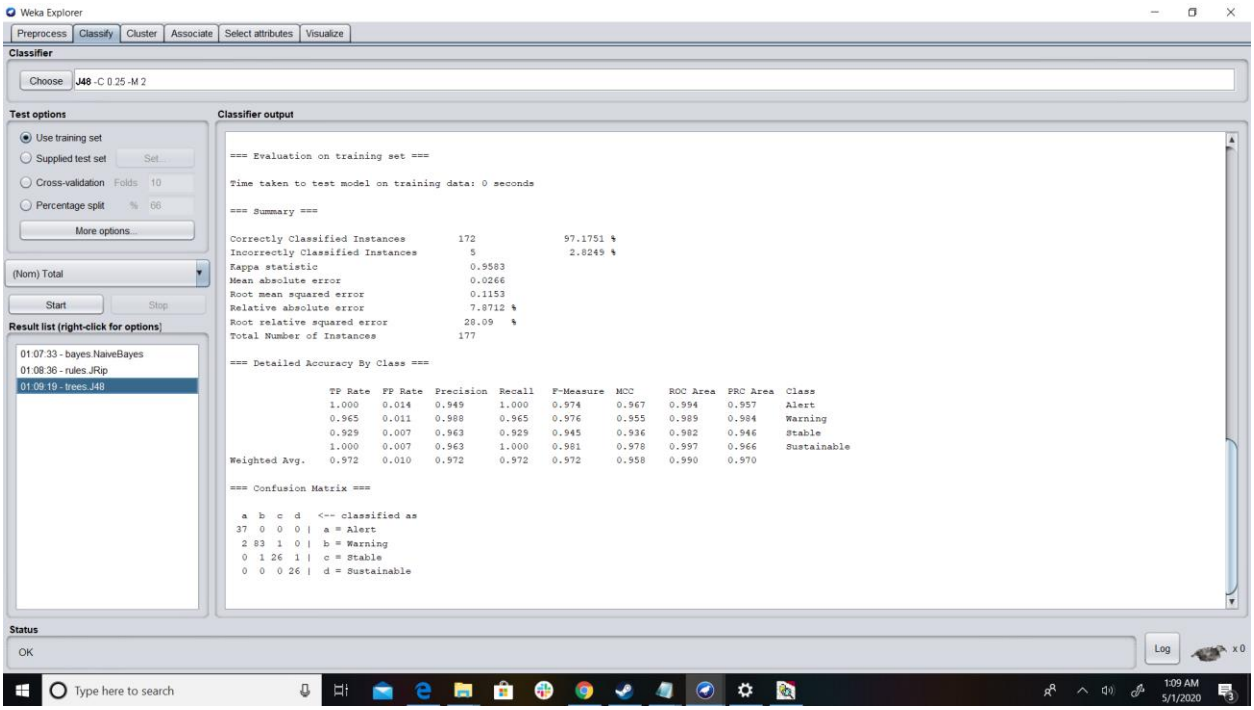
1. Naïve Bayes:



2. JRIP


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
JRip -F 3 -N 2.0 -O 2 -S 1
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
(Nom) Total
Start
Stop
Result list (right-click for options)
01:07:33 - bayes.NaiveBayes
01:08:36 - rules.JRip
Classifier output
=== Evaluation on training set ===
Time taken to test model on training data: 0 seconds
=== Summary ===
Correctly Classified Instances 176 95.435 %
Incorrectly Classified Instances 1 0.565 %
Kappa statistic 0.9916
Mean absolute error 0.0054
Root mean squared error 0.0521
Relative absolute error 1.6063 %
Root relative squared error 12.6895 %
Total Number of Instances 177
=== 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 Alert
1.000 0.000 1.000 1.000 1.000 1.000 1.000 1.000 Warning
0.964 0.000 1.000 0.964 0.982 0.979 0.997 0.984 Stable
1.000 0.007 0.963 1.000 0.981 0.978 0.997 0.966 Sustainable
Weighted Avg. 0.994 0.001 0.995 0.994 0.994 0.993 0.999 0.992
=== Confusion Matrix ===
a b c d <-- classified as
37 0 0 0 | a = Alert
0 86 0 0 | b = Warning
0 0 27 1 | c = Stable
0 0 0 26 | d = Sustainable
Status
OK
Log
x 0
Type here to search
1:08 AM
5/1/2020

3. JR8


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) Total
Start
Stop
Result list (right-click for options)
01:07:33 - bayes.NaiveBayes
01:08:36 - rules.JRip
01:09:19 - trees.J48
Classifier output
=== Evaluation on training set ===
Time taken to test model on training data: 0 seconds
=== Summary ===
Correctly Classified Instances 172 97.1751 %
Incorrectly Classified Instances 5 2.8249 %
Kappa statistic 0.9583
Mean absolute error 0.0266
Root mean squared error 0.1153
Relative absolute error 7.8712 %
Root relative squared error 28.09 %
Total Number of Instances 177
=== Detailed Accuracy By Class ===
TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class
1.000 0.014 0.949 1.000 0.974 0.967 0.994 0.957 Alert
0.965 0.011 0.988 0.965 0.976 0.955 0.989 0.984 Warning
0.929 0.007 0.963 0.929 0.945 0.936 0.982 0.946 Stable
1.000 0.007 0.963 1.000 0.981 0.978 0.997 0.966 Sustainable
Weighted Avg. 0.972 0.010 0.972 0.972 0.972 0.958 0.990 0.970
=== Confusion Matrix ===
a b c d <-- classified as
37 0 0 0 | a = Alert
2 83 1 0 | b = Warning
0 1 26 1 | c = Stable
0 0 0 26 | d = Sustainable
Status
OK
Log
x 0
Type here to search
1:09 AM
5/1/2020

V. Generation of Action Rules using LISP Miner

Attributes:

We have classified our attributes into 3 categories

1. Stable
2. Flexible
3. Decision

Stable Attributes:

1. Country
2. Rank
3. Year

Flexible Attributes:

1. Exports and Imports
2. Security Apparatus
3. Economy
4. State Legitimacy
5. Life Expectancy at Birth
6. Demographic Pressures
7. Unemployment Total

Decision Attributes(Total)

Category

1. Sustainable
2. Stable
3. Warning
4. Alert

Antecedent Stable Part:

We assign all the stable attributes to this set.

Antecedent Variable Part:

We assign all the flexible attributes to this set.

Succedent Variable Part:

We assign the decision variable to this set.

Attribute type = Nominal

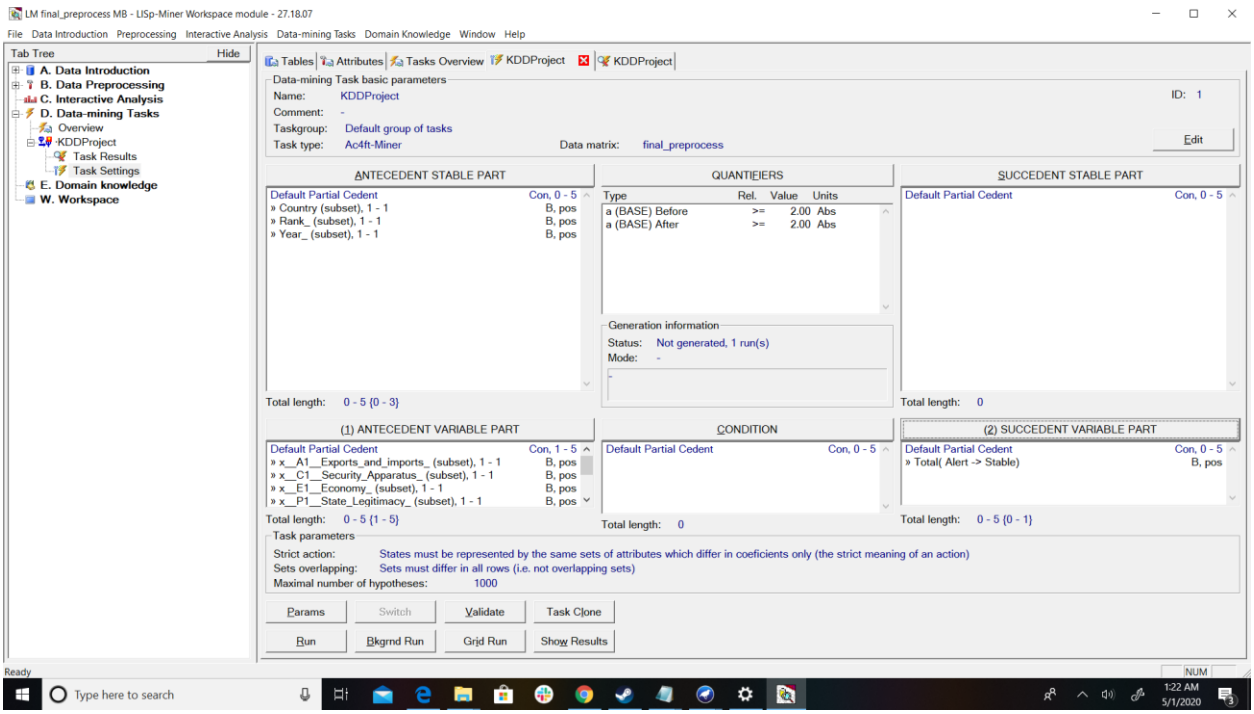
Coefficient type = one category

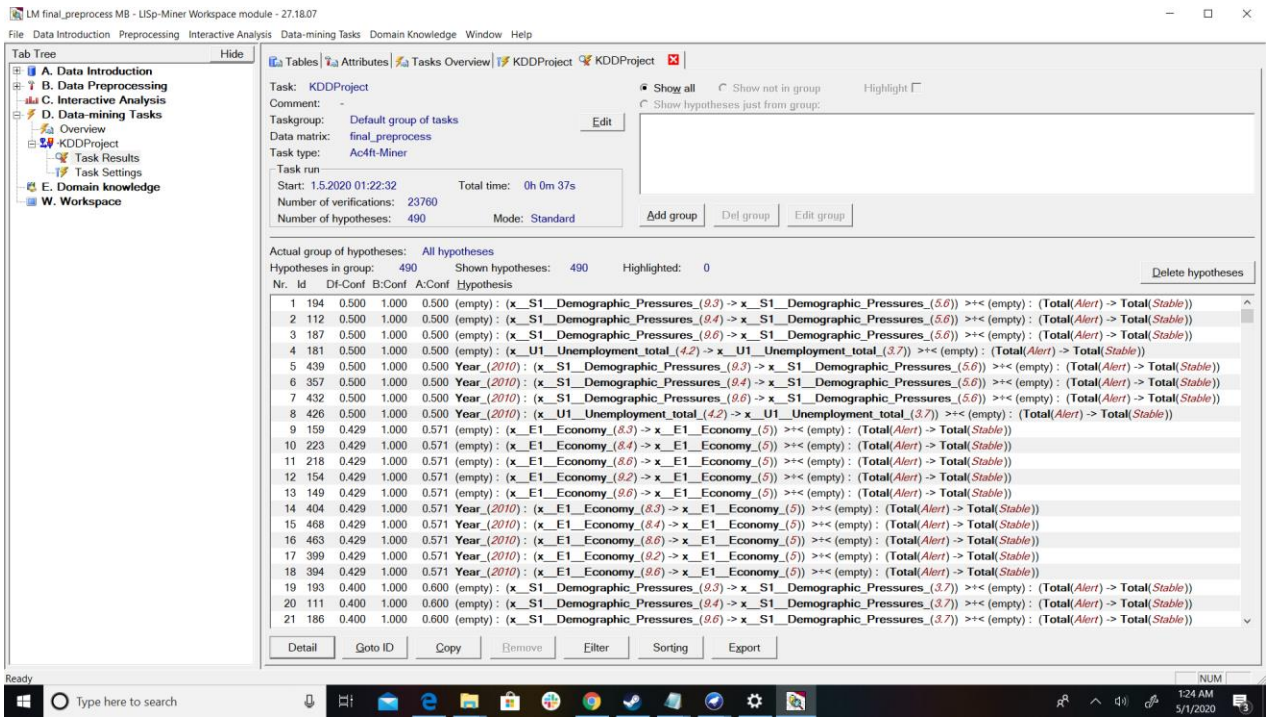
Quantifiers:

A(BASE)Before: 2

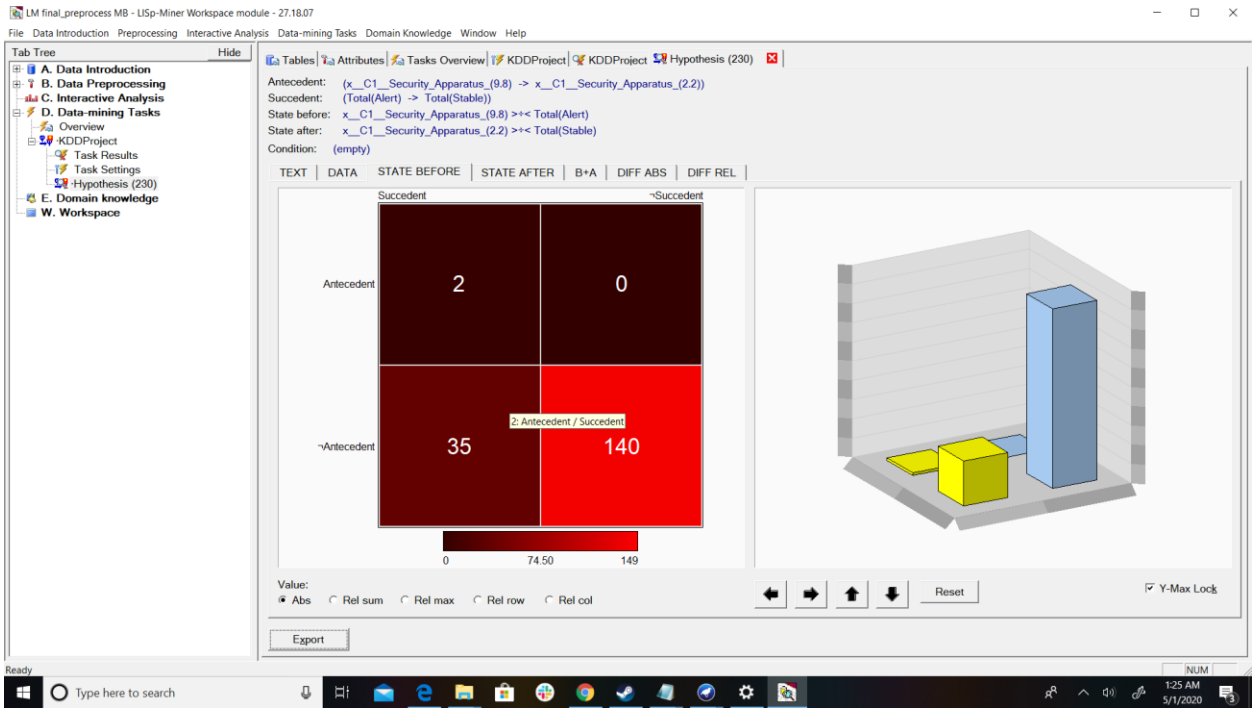
A(BASE)After : 2

Action Rules Alert-> Stable

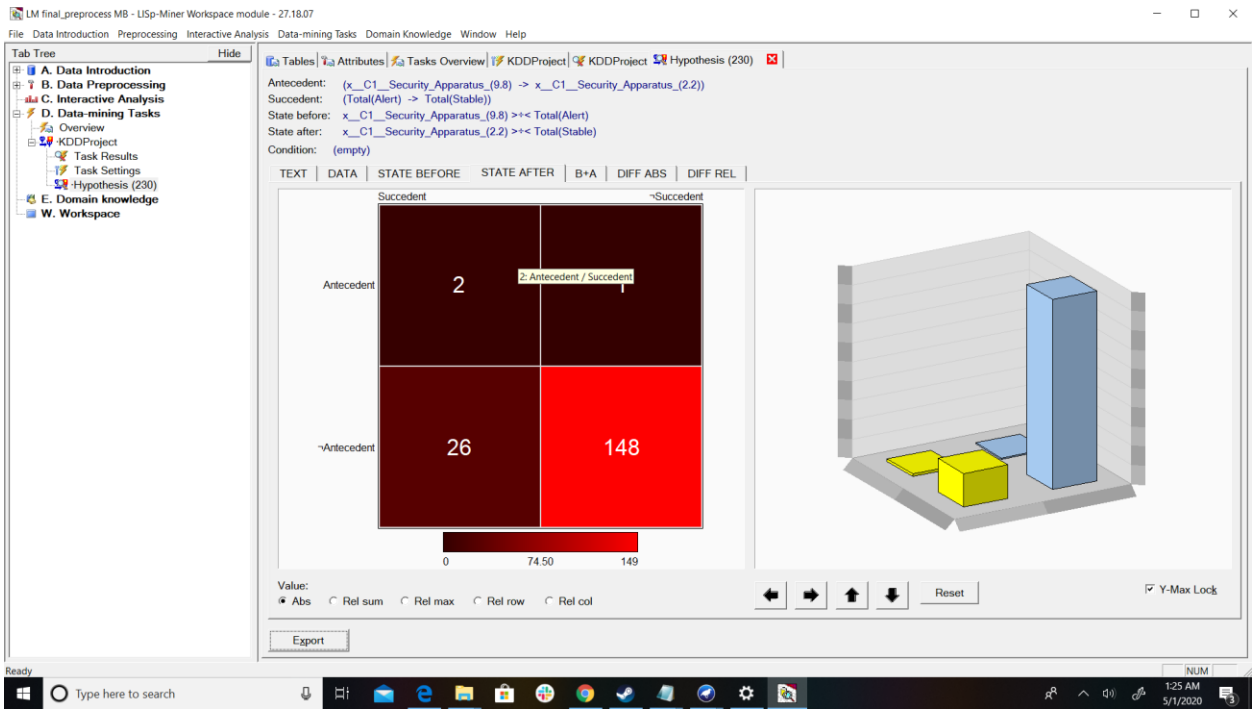




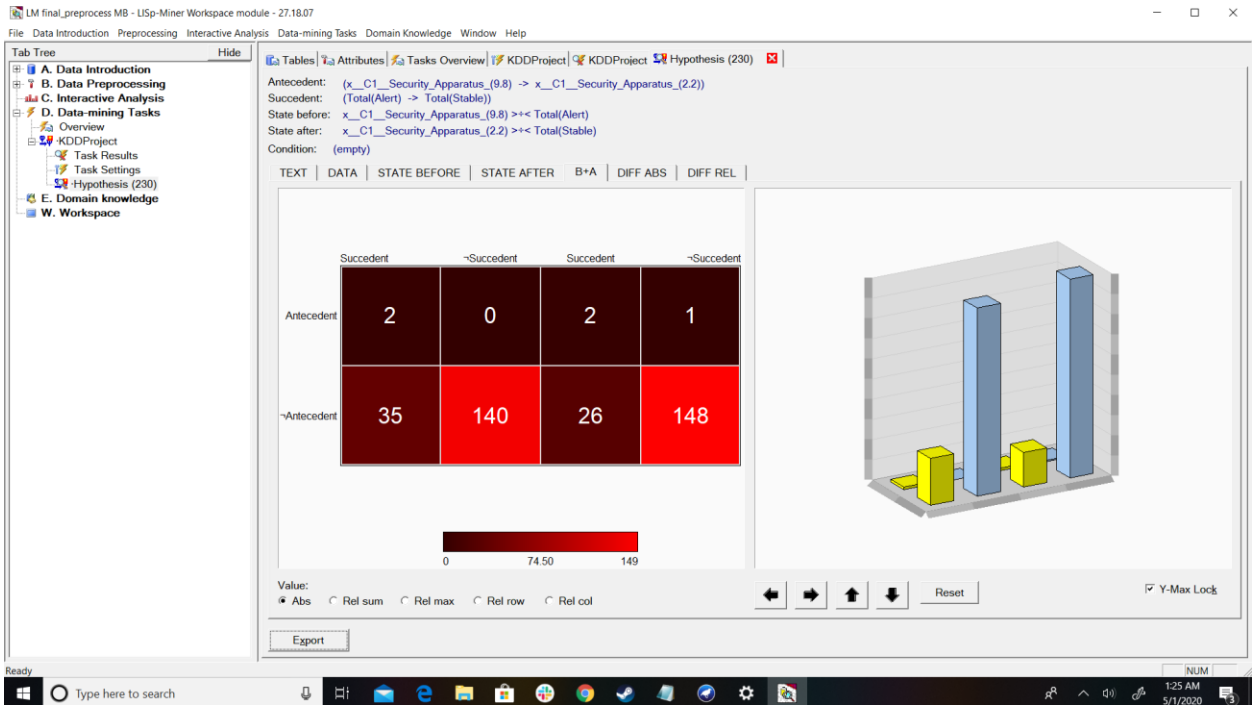
Hypothesis and State Before:



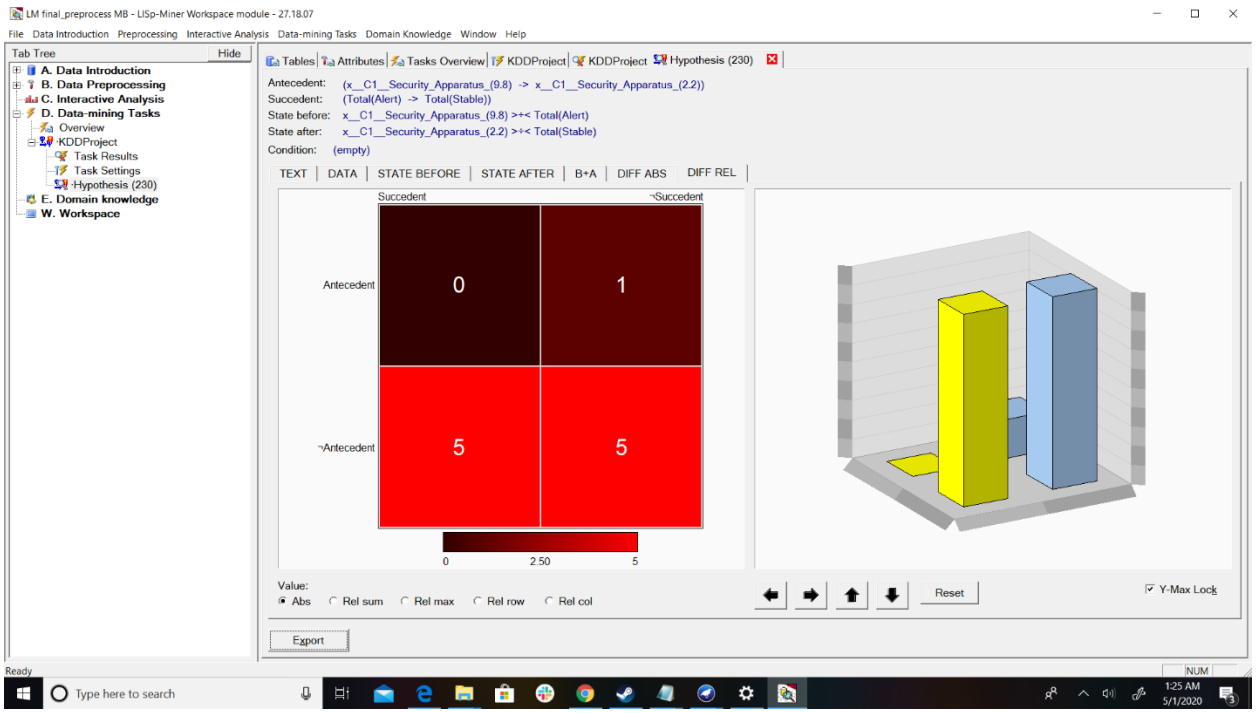
Hypothesis and State After:



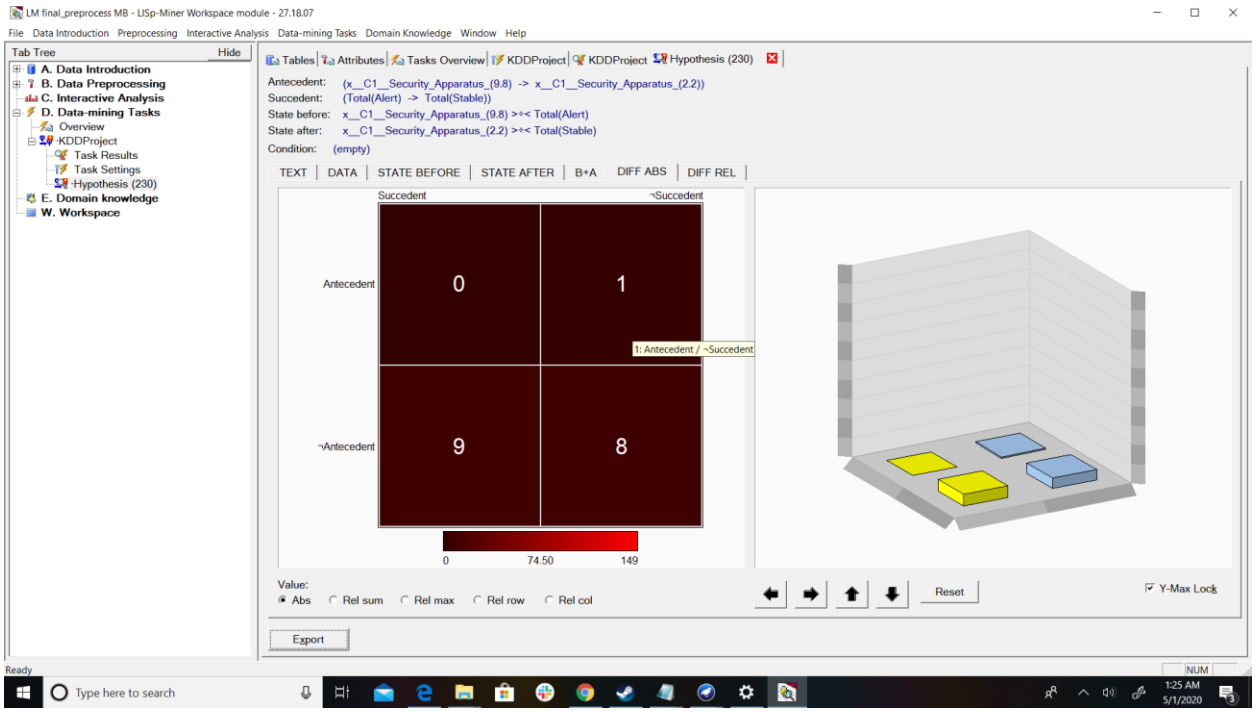
B+A:



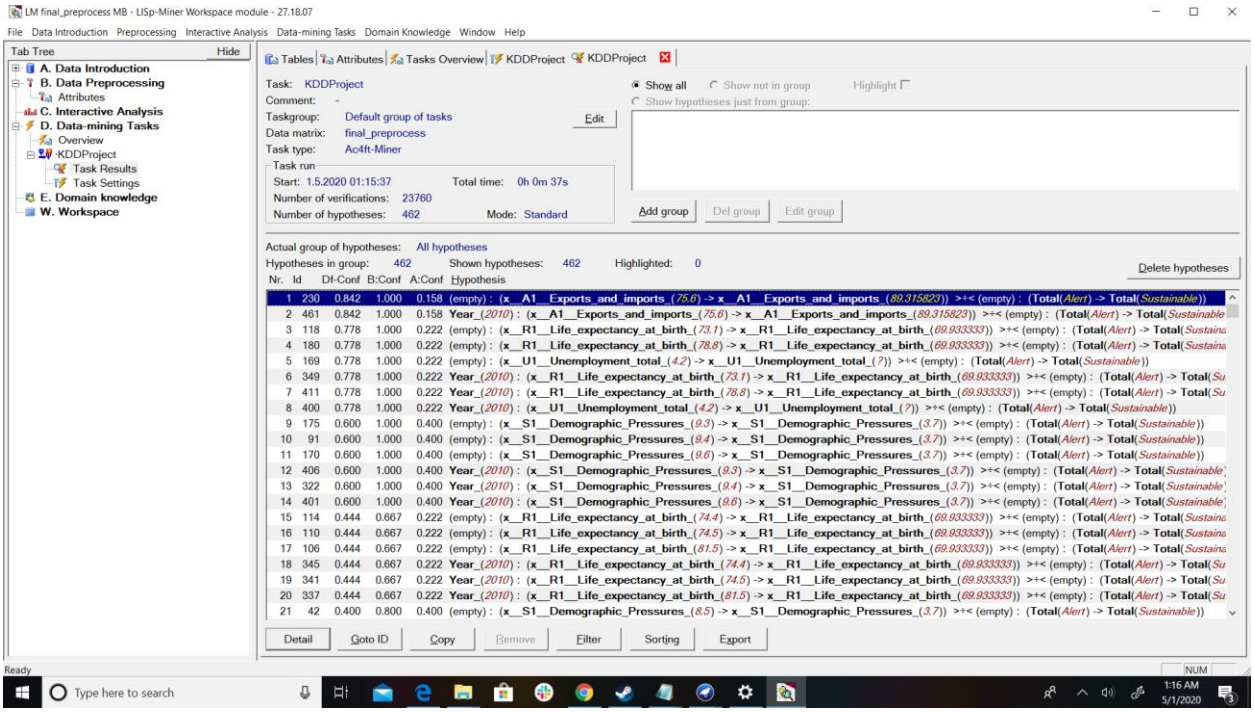
Diff Rel:

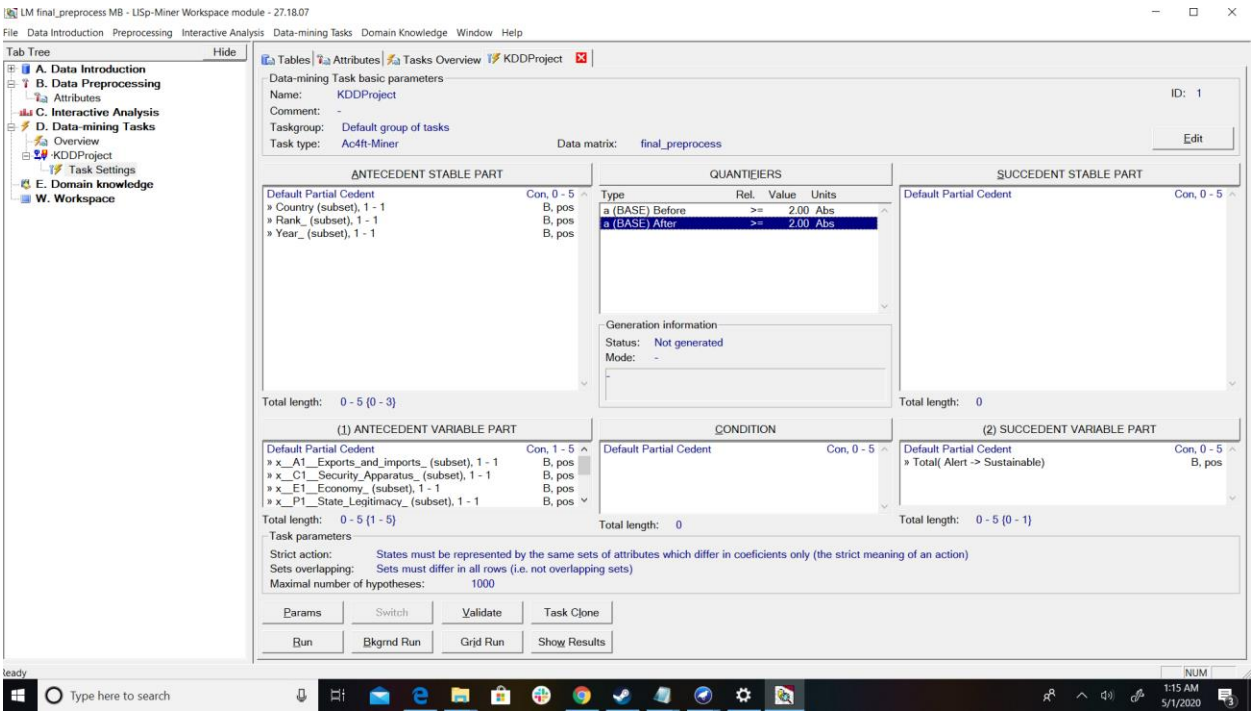


Diff Abs:

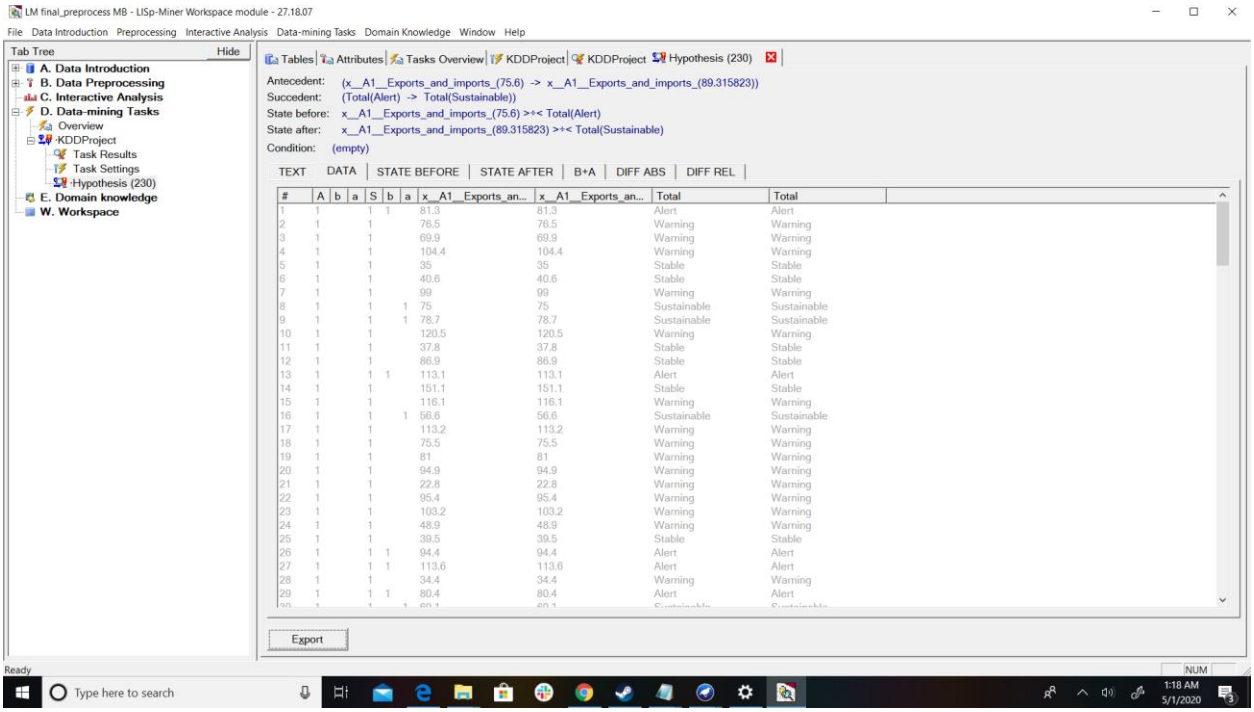


Alert ->Sustainable

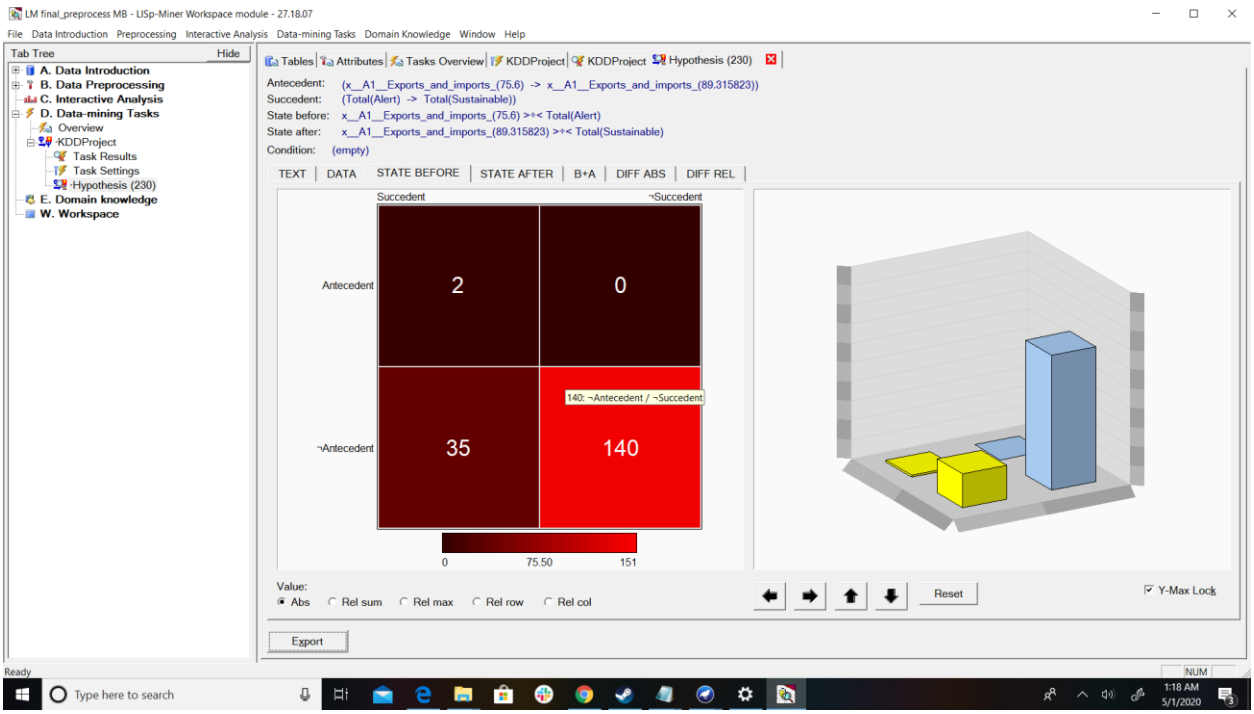




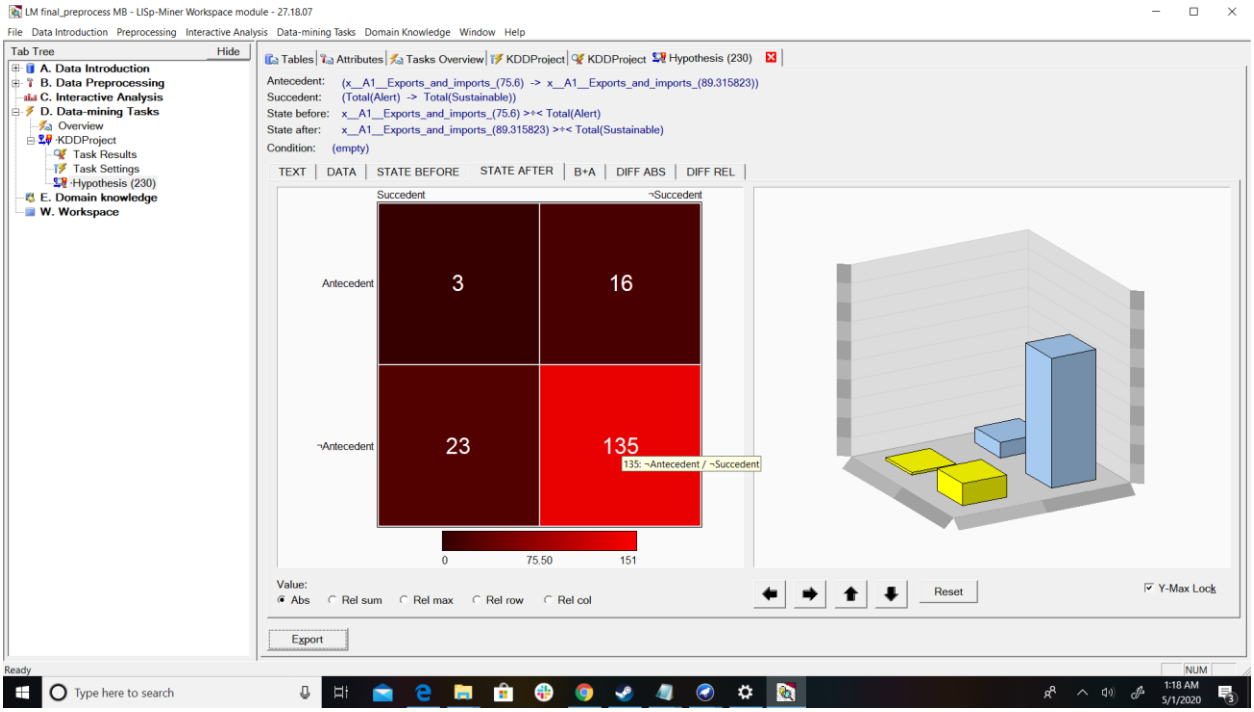
Data:



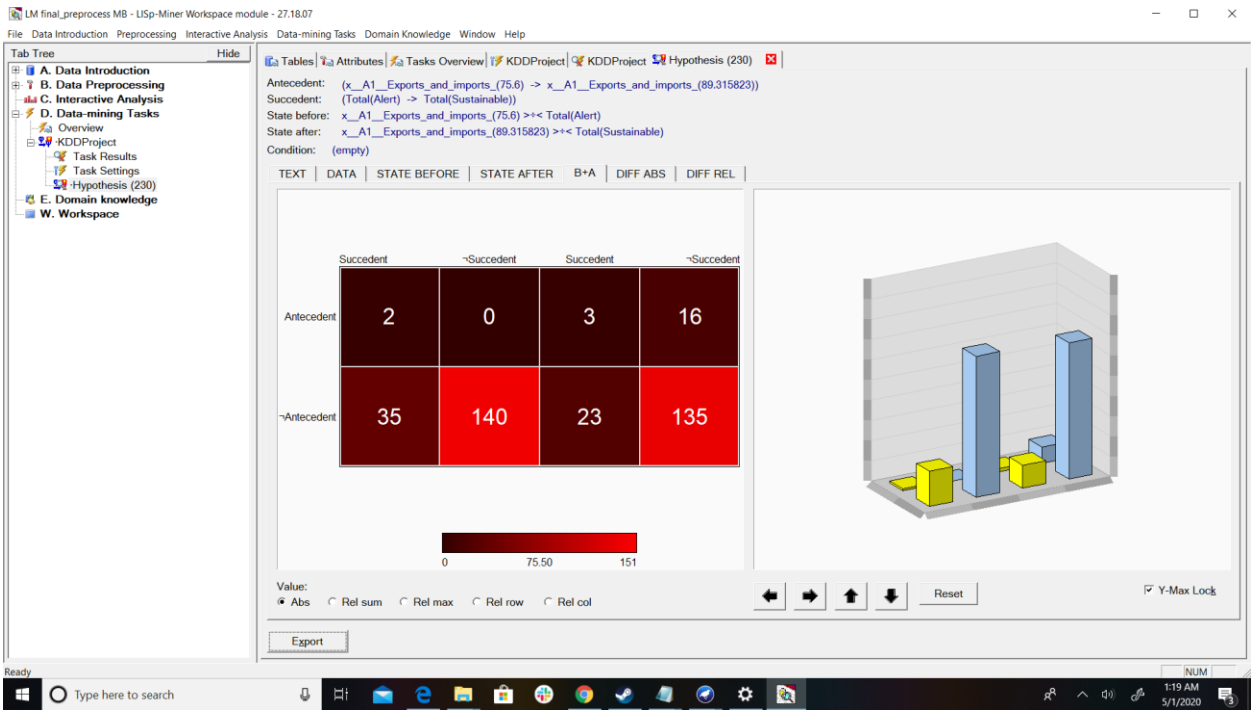
State Before:



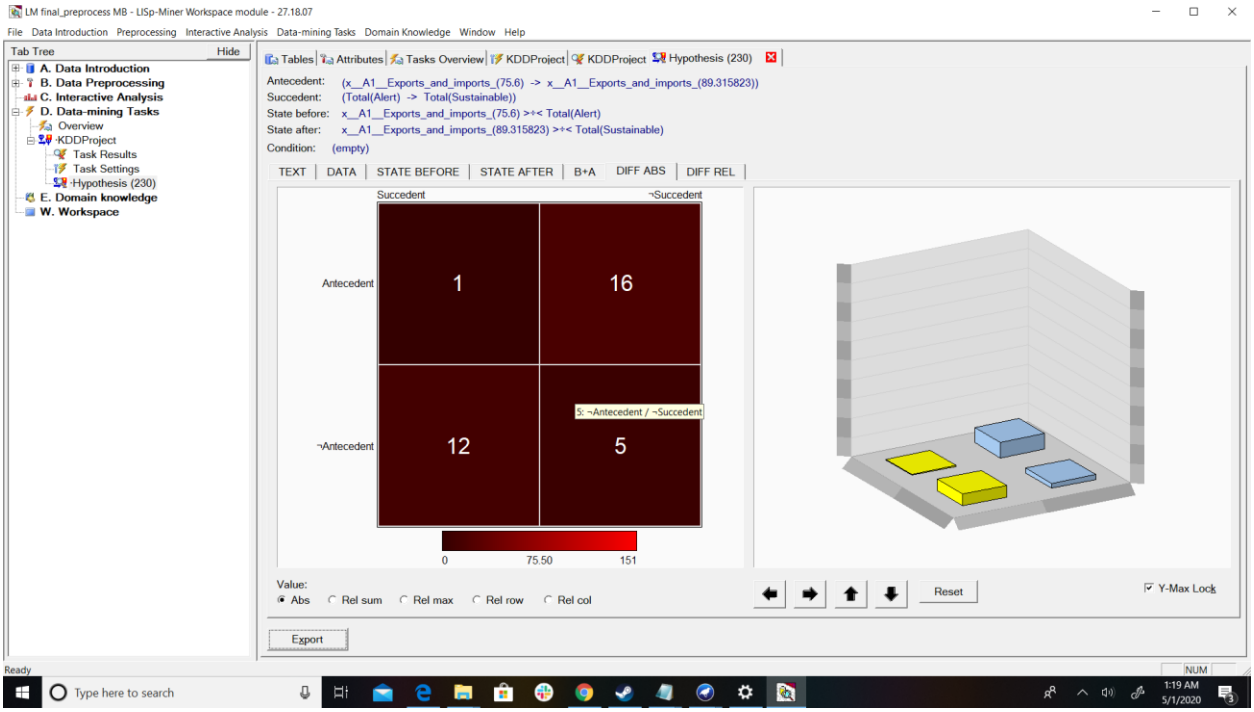
State After:



B+A



Diff ABS:



Diff Rel: