Movie Ratings Analysis: Predictions and Strategies for 9,000 Films

ITCS 6162

KNOWLEDGE DISCOVERY IN DATABASES

- by DR. ZBIGNIEW W RAS

By Group 14 (Name: CipherNova)

Hamsini Battula (801328682)

Pooja Haridasyam (801328954)

Sreeja Chevula (801329242)

Jahnavi Muddu (801329244)

PROJECT DESCRIPTION

Analyzing a selection of 9,000 movies from the Movies Database is the main focus of the problem. Understanding and forecasting consumer ratings for these films is the main goal. Our objective is to compute average ratings using user data from the "ratings.small" file in order to accomplish this. The project requires proposing and adding four new classification qualities to the decision table in addition to the ones that are already there. Customer ratings are predicted to be connected with these variables, which could include cast popularity, release year, director repute, and popularity of the genre.

Developing efficient classifiers using technologies like Orange or Weka is the main problem. The comparison uses the F-score measure to assess classifier performance both with and without the recently added characteristics. A classifier with better performance when considering the four more attributes is the expected result.

Moreover, the research extends to Lisp-Miner with the goal of obtaining action rules that provide practical suggestions for improving movie ratings. These guidelines will shed light on particular acts that might raise the box office receipts of particular films.

To put it simply, the issue domain includes feature engineering, data extraction, classifier construction, and rule mining, all of which are intended to provide a more comprehensive grasp of the variables affecting customer movie evaluations.

WHAT ARE THE NEW FEATURES?

In addition to the initial CSV file, this report incorporates four additional features. Below, we provide a list of these four newly introduced features, accompanied by the rationale behind their selection.

1. **SeasonReleased:** It is derived from the 'release_date' column and shows the month (as an integer) that the film was released. By giving the release month a categorical representation, it makes seasonality-based analysis possible. For instance, various popularity trends may be seen for films that are released during particular seasons or holidays.

2. **simplifiedPopularity:** It is the streamlined and rounded integer depiction of the film's popularity, derived from the 'popularity' column. Provides a clearer picture of the film's popularity by averaging the initial "popularity" score. Movies can be more easily interpreted by using this simplification to group them into popularity ranges.

3. **Profitability_numeric:** Produced by assigning the three labels of -1 (negative profit), 0 (break-even), and 1 (positive profit) to the "profit" (revenue - budget). gives the financial health of a film in numerical form. Films are categorized according to their financial success or failure, which streamlines the study.

4. **Popularity_score:** Calculated by multiplying "vote_count" by "vote_average," then standardizing the outcome. Combines the number of votes ('vote_count') and the average vote ('vote_average') to get a weighted popularity score. By ensuring a consistent measurement, normalization provides a thorough understanding of a film's popularity by accounting for both the amount and caliber of votes.

5. **Average_rating:** 'average_rating' represents the mean score given by viewers, calculated from individual ratings. It provides a concise measure of a movie's overall reception and is crucial for assessing audience satisfaction and predicting a film's success in the industry.

WHY WE CHOOSE THESE FEATURES?

Based on the desire to improve the dataset for a more thorough study of films, the four features in the script were chosen. Let's examine the reasoning behind choosing each feature in more detail.

1. **SeasonReleased:** A movie's performance can be greatly impacted by the season of its release. The popularity and financial success of films that are released around holidays or during particular seasons may be affected by the different audience they draw.
2. **SimplifiedPopularity:** The popularity score can be made more easily comprehensible and categorized by simplifying it. This balanced portrayal makes it simpler to compare and arrange films according to their general level of popularity.
3. **Profitability_numeric:** Sorting movies according to their profitability levels offers a clear-cut way to determine if they are profitable or not. For stakeholders, such producers and investors, to swiftly evaluate the financial performance of a film, this might be essential.
4. **Popularity_score:** A more sophisticated gauge of a film's general appeal is offered by the weighted popularity score, which adds together the total number of votes and the average vote. This aids in comprehending the caliber of the audience's opinions in addition to the number of votes.
5. **Average_rating:** The "average_rating" attribute was chosen for its role as a consolidated metric, offering a comprehensive evaluation of a movie's reception. Widely recognized in the industry, it serves as an informative benchmark for decision-making in production, marketing, and distribution.

METHOD AND SOURCE OF EXTRACTION

Using Google Colab, the four properties were extracted from the movies_metadata.csv. We created the fundamental dataset by combining links_small.csv and ratings_small.csv using an inner join on tmdbid. The four additional attributes were then integrated into the original file to enrich it and create an expanded dataset. To find important rules and insights, this expanded dataset was subjected to additional analysis, specifically action rule mining.

Extraction of the attributes:

1. **SeasonReleased:** SeasonReleased is a numeric indication of a film's release month. obtained by taking the'release_date', formatting it in datetime, then extracting the month as a number.

2. **simplifiedPopularity:** A movie's popularity score expressed as a rounded, simplified integer. determined by converting the "popularity" value to a numeric representation and then rounding it.

3. **profitability_numeric:** Movies are categorized by profitability using the profitability_numeric function, where a value of -1 denotes a loss, 0 represents a break-even point, and 1 represents a profit. Ascertained by assigning three labels (-1, 0 and 1) to the 'profit' (revenue - budget).

4. **popularity_score:** A popularity value that is weighted and considers the average and total number of votes cast. Vote_count and Vote_average is multiplied, the result is normalized, and it is rounded to two decimal points.

DATA PREPROCESSING

When working with large datasets from many sources, data preparation becomes essential since these datasets frequently contain irrelevant information that could skew the results of later analysis. Unwanted elements like features and null values are eliminated during this preprocessing stage.

Data Transformation:

The NumericToNominal procedure, which transforms numerical data into categorical data, is an essential part of data transformation. For datasets with a limited number of possible values, like ratings, survey answers, or grades, this translation works especially well. Each numerical value is given a categorical label by NumericToNominal, making it simpler to analyze and understand the data. Moreover, the data is transformed to prepare it for particular machine learning algorithms that require categorical input.

CLASSIFIERS USING WEKA ON ORIGINAL DATASET

Sorting data into distinct groups in accordance with predetermined standards is the process of classification. Whether handling large or unstructured information, the classification procedure entails examining and classifying the data into discrete groups. This makes it easier to determine which category new data belongs in, which streamlines the integration process.

The goal here is achieved by the project using five distinct categorization algorithms.

In the dataset, we transformed the average_rating attribute from a numerical format to a categorical one. The illustration below depicts this conversion process.

Loading Dataset into WEKA

Fig: Launch the Weka software and proceed by choosing the explorer option.

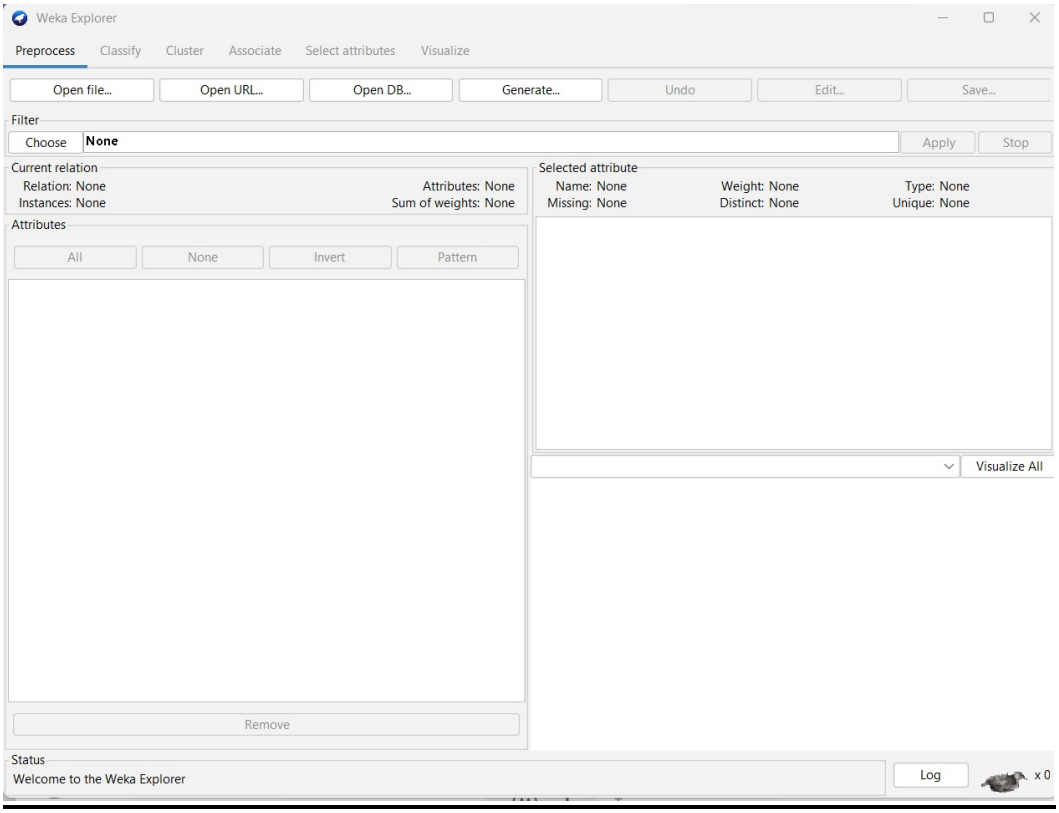
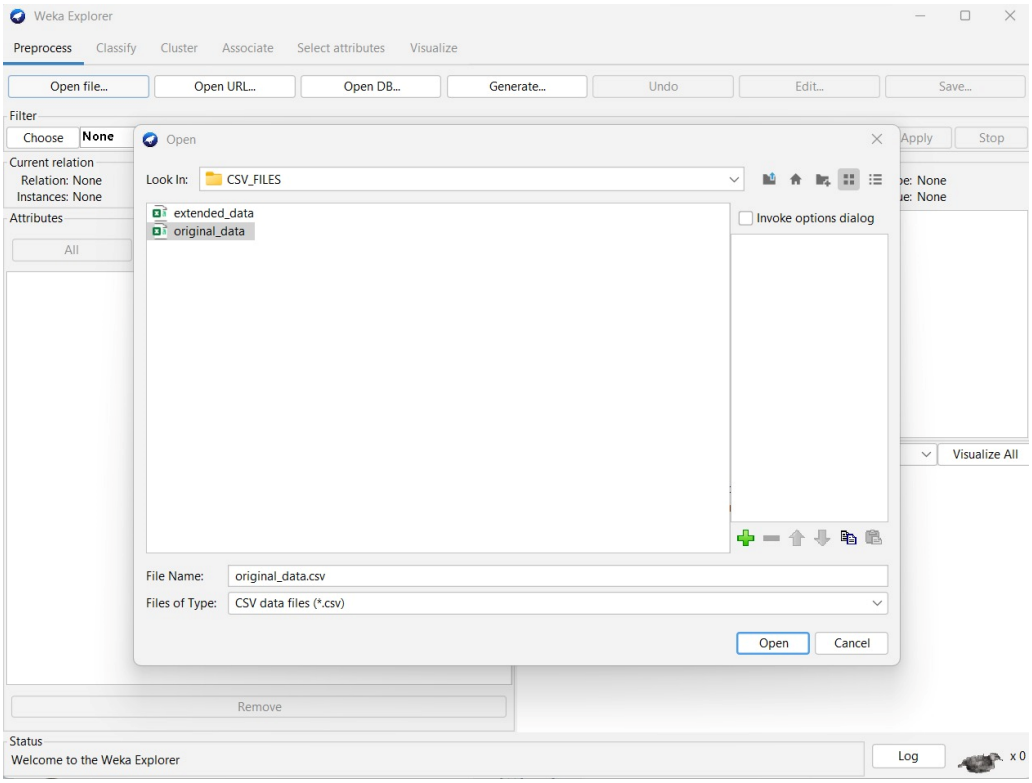


Fig: choose the "Open File" option to import a CSV file.



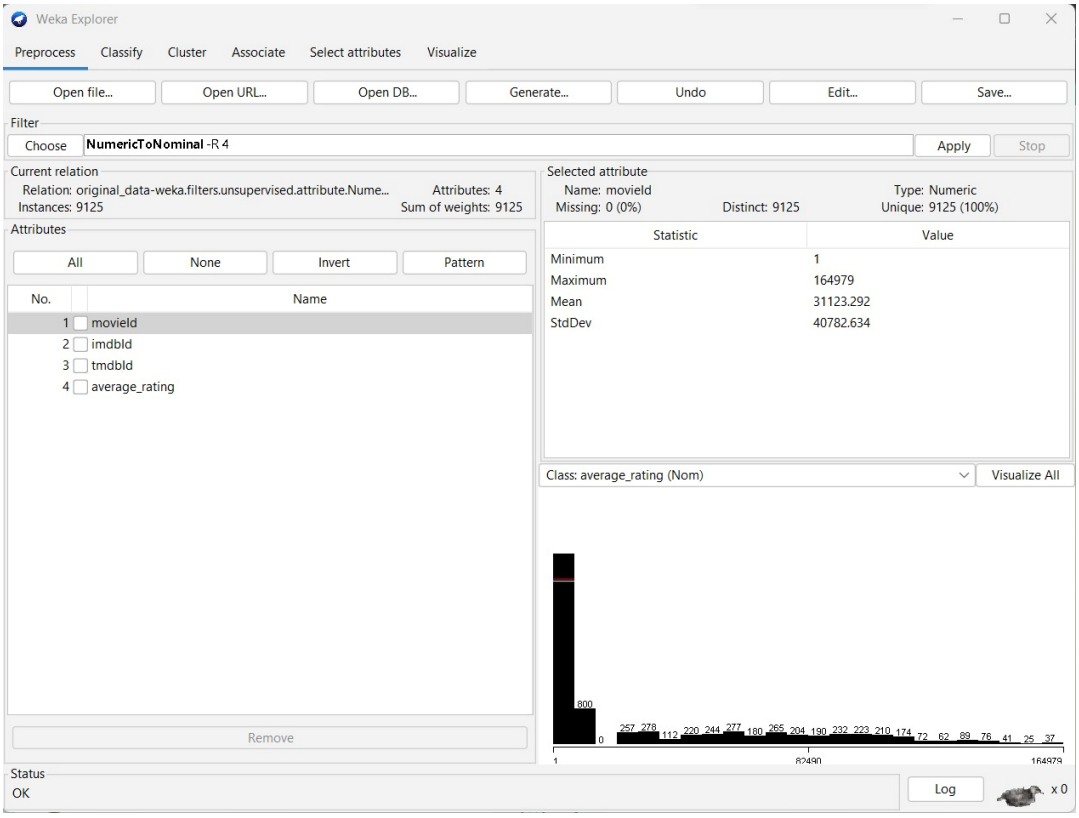


Fig: Select the filter from the options below, and then apply the specified actions to the data.

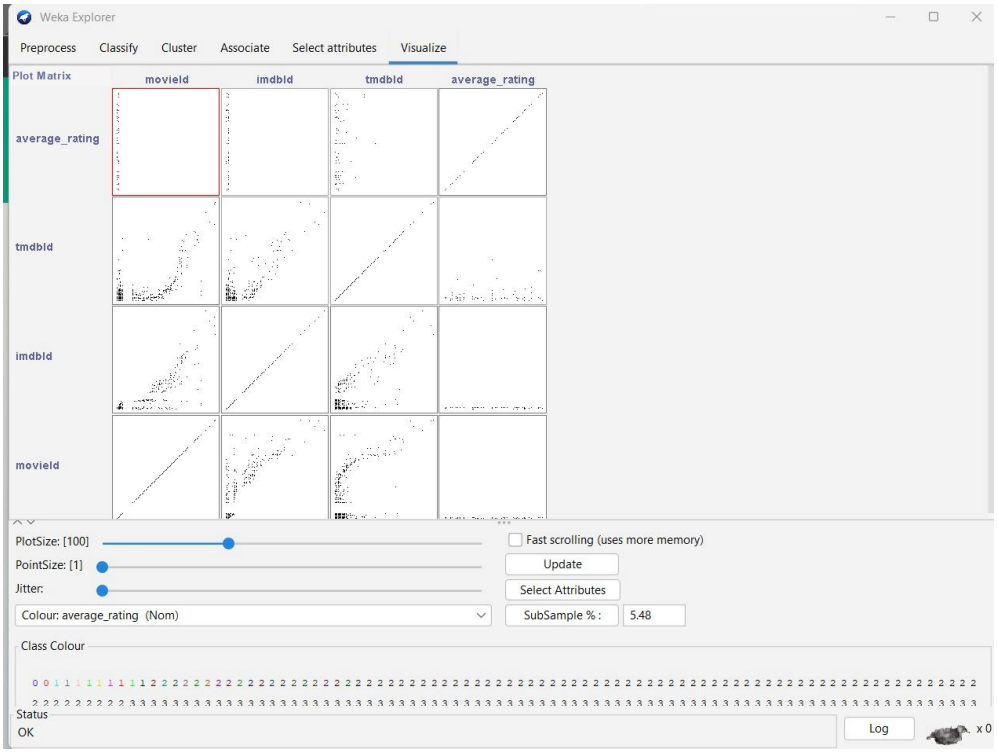


Fig: After clicking the visualize button, the outcomes are displayed.

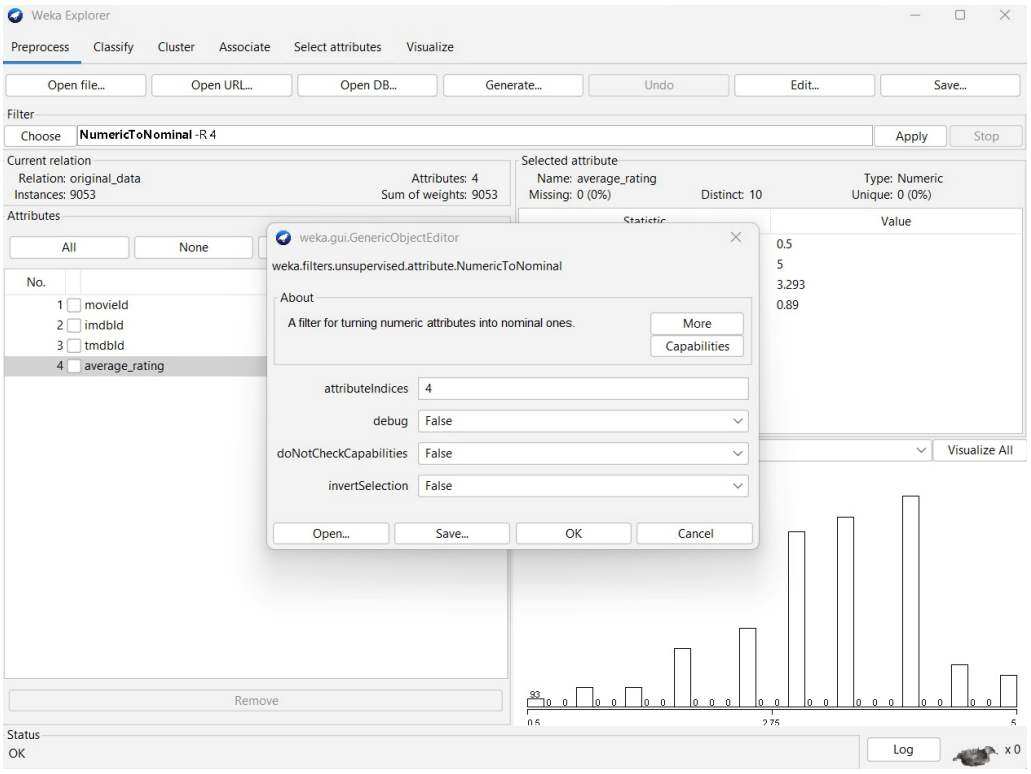
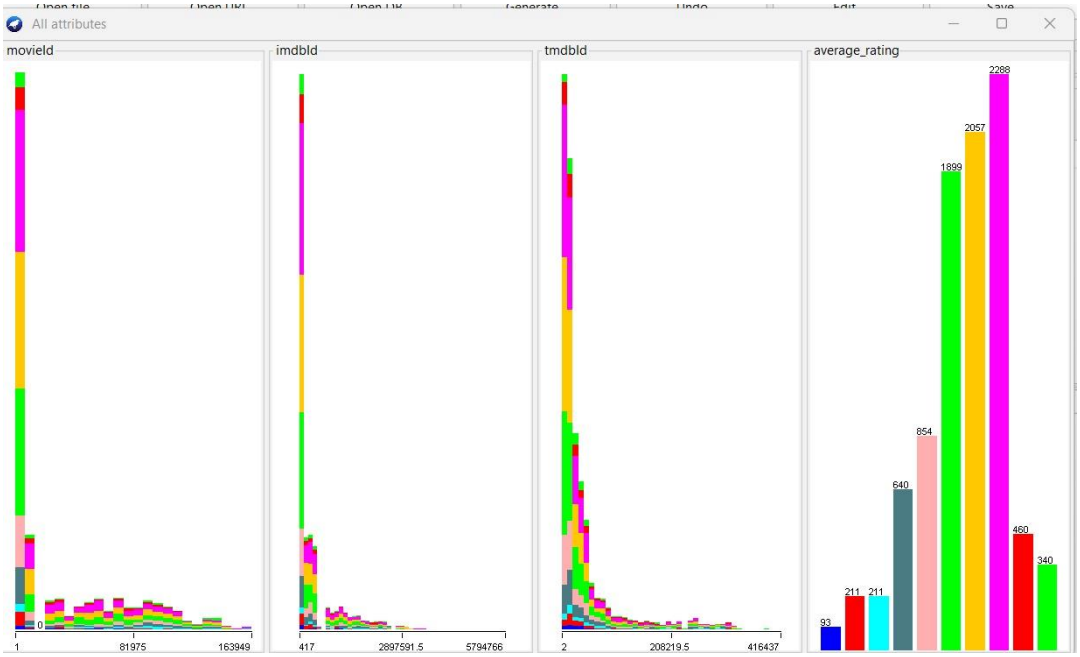


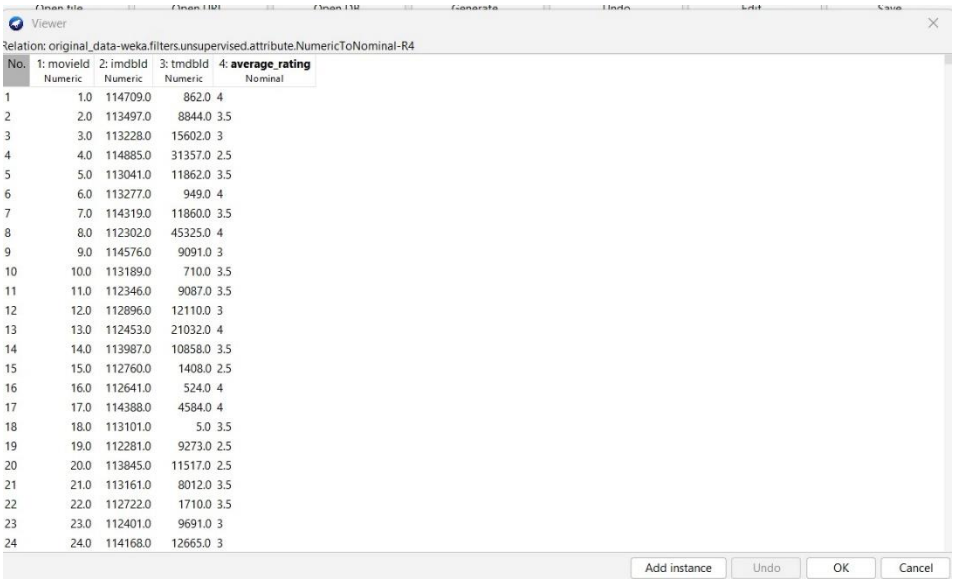
Fig: Loading dataset into WEKA

Upon selecting the dataset, we conducted a data transformation on the fourth column, namely average_rating. This process involves converting the numerical values of average_rating into nominal values.

Visualisation:



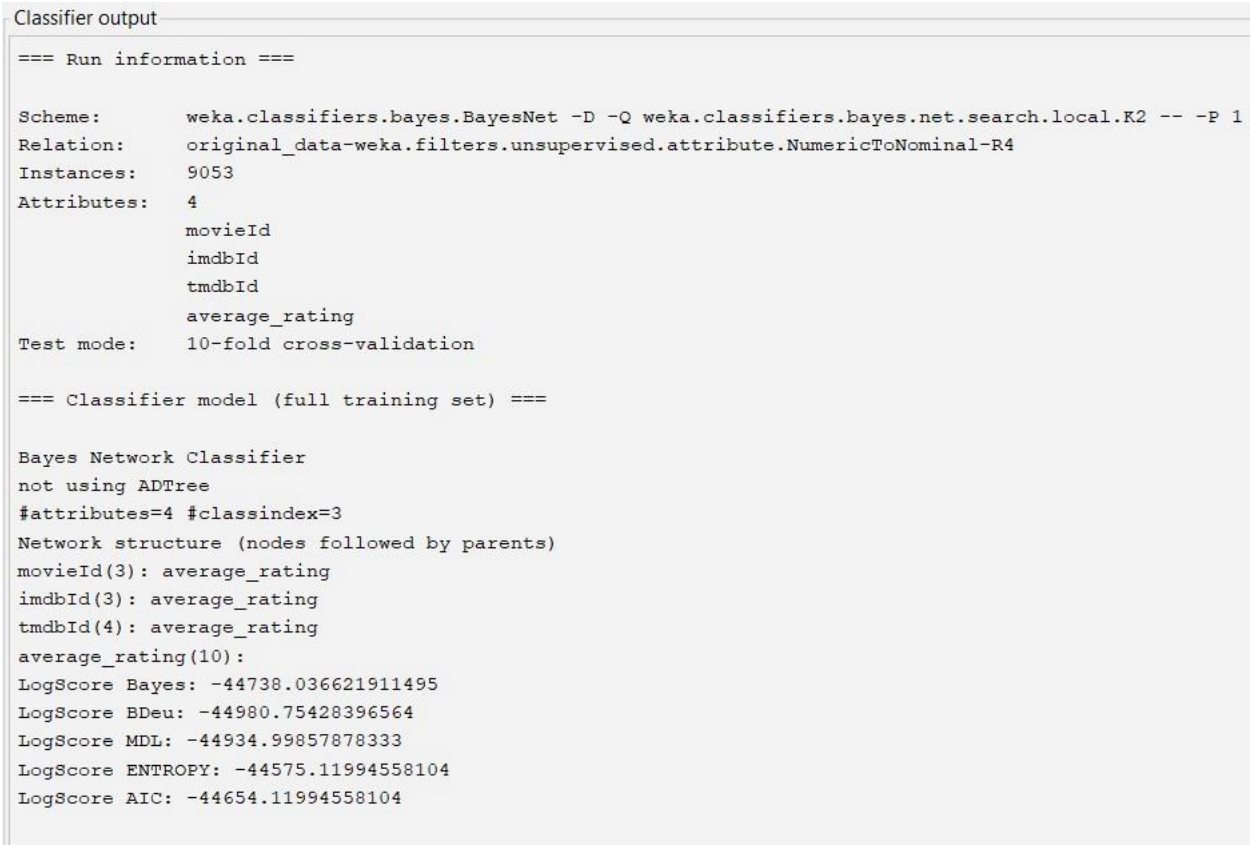
By clicking on “Edit” we’ve obtained the preprocessed data.



No.	1: movieId Numeric	2: imdbId Numeric	3: tmdbId Numeric	4: **average_rating** Nominal
1	1.0	114709.0	862.0	4
2	2.0	113497.0	8844.0	3.5
3	3.0	113228.0	15602.0	3
4	4.0	114885.0	31357.0	2.5
5	5.0	113041.0	11862.0	3.5
6	6.0	113277.0	949.0	4
7	7.0	114319.0	11860.0	3.5
8	8.0	112302.0	45325.0	4
9	9.0	114576.0	9091.0	3
10	10.0	113189.0	710.0	3.5
11	11.0	112346.0	9087.0	3.5
12	12.0	112896.0	12110.0	3
13	13.0	112453.0	21032.0	4
14	14.0	113987.0	10858.0	3.5
15	15.0	112760.0	1408.0	2.5
16	16.0	112641.0	524.0	4
17	17.0	114388.0	4584.0	4
18	18.0	113101.0	5.0	3.5
19	19.0	112281.0	9273.0	2.5
20	20.0	113845.0	11517.0	2.5
21	21.0	113161.0	8012.0	3.5
22	22.0	112722.0	1710.0	3.5
23	23.0	112401.0	9691.0	3
24	24.0	114168.0	12665.0	3

The dataset has been appropriately processed and is now composed for classification analysis.

1. **BAYESNET CLASSIFICATION:** The probability of a discrete class variable C, given specific features X, is predicted by the Simple Bayesian Network Classifier and is represented as P(c|x). This technique finds the best-fitting Bayesian network in a dataset D over U by using search strategies and quality metrics. The main goal of the learning challenge is to identify the best Bayesian network that fits the features of the provided dataset.



```
Time taken to build model: 0 seconds

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

Correctly Classified Instances        2331               25.7484 %
Incorrectly Classified Instances      6722               74.2516 %
Kappa statistic                          0.0538
Mean absolute error                      0.1619
Root mean squared error                  0.2866
Relative absolute error                 98.5507 %
Root relative squared error            100.008  %
Total Number of Instances             9053

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.000    0.000    ?          0.000    ?          ?        0.700     0.018     0.5
                 0.000    0.000    ?          0.000    ?          ?        0.614     0.032     1
                 0.346    0.099    0.077      0.346    0.126      0.122    0.706     0.052     1.5
                 0.003    0.001    0.182      0.003    0.006      0.015    0.599     0.093     2
                 0.000    0.000    ?          0.000    ?          ?        0.561     0.107     2.5
                 0.175    0.157    0.229      0.175    0.198      0.020    0.559     0.244     3
                 0.312    0.235    0.281      0.312    0.296      0.075    0.579     0.278     3.5
                 0.560    0.455    0.294      0.560    0.386      0.091    0.577     0.324     4
                 0.000    0.000    ?          0.000    ?          ?        0.610     0.067     4.5
                 0.000    0.000    ?          0.000    ?          ?        0.664     0.061     5
Weighted Avg.    0.257    0.204    ?          0.257    ?          ?        0.584     0.221
```

```
=== Confusion Matrix ===

    a    b    c    d    e    f    g    h    i    j   <-- classified as
    0    0    7    0    0   15    3   68    0    0 |    a = 0.5
    0    0   25    0    0   37   43  106    0    0 |    b = 1
    0    0   73    0    0   26   29   83    0    0 |    c = 1.5
    0    0  116    2    0  133  144  245    0    0 |    d = 2
    0    0  101    0    0  172  253  328    0    0 |    e = 2.5
    0    0  176    3    0  332  582  806    0    0 |    f = 3
    0    0  153    3    0  331  642  928    0    0 |    g = 3.5
    0    0  197    0    0  328  481 1282    0    0 |    h = 4
    0    0   52    0    0   49   66  293    0    0 |    i = 4.5
    0    0   45    3    0   29   41  222    0    0 |    j = 5
```

=== Run information ===

Scheme: weka.classifiers.bayes.BayesNet -D -Q weka.classifiers.bayes.net.search.local.K2 -- -P 1 -S BAYES -E weka.classifiers.bayes.net.estimate.SimpleEstimator -- -A 0.5

Relation: original_data-weka.filters.unsupervised.attribute.NumericToNominal-R4

Instances: 9053

Attributes: 4

movieId

imdbId

tmdbId

average_rating

Test mode: 10-fold cross-validation

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

Bayes Network Classifier

not using ADTree

#attributes=4 #classindex=3

Network structure (nodes followed by parents)

movieId(3): average_rating

imdbId(3): average_rating

tmdbId(4): average_rating

average_rating(10):

LogScore Bayes: -44738.036621911495

LogScore BDeu: -44980.75428396564

LogScore MDL: -44934.99857878333

LogScore ENTROPY: -44575.11994558104

LogScore AIC: -44654.11994558104

Time taken to build model: 0 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	2331	25.7484 %
Incorrectly Classified Instances	6722	74.2516 %
Kappa statistic	0.0538	
Mean absolute error	0.1619	
Root mean squared error	0.2866	
Relative absolute error	98.5507 %	
Root relative squared error	100.008 %	
Total Number of Instances	9053	

=== Detailed Accuracy By Class ===

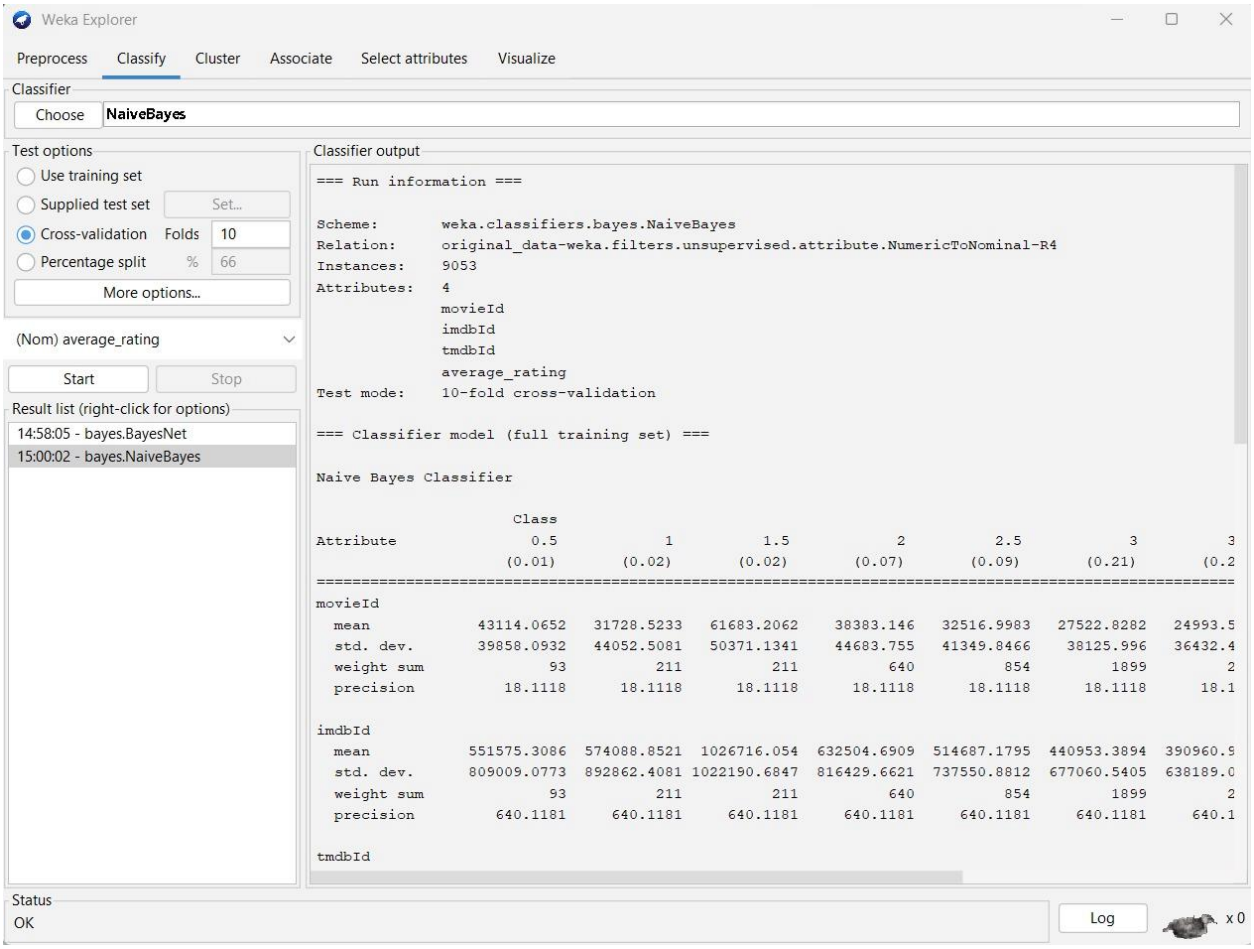
TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0.000	0.000	?	0.000	?	?	0.700	0.018	0.5
0.000	0.000	?	0.000	?	?	0.614	0.032	1
0.346	0.099	0.077	0.346	0.126	0.122	0.706	0.052	1.5
0.003	0.001	0.182	0.003	0.006	0.015	0.599	0.093	2
0.000	0.000	?	0.000	?	?	0.561	0.107	2.5
0.175	0.157	0.229	0.175	0.198	0.020	0.559	0.244	3
0.312	0.235	0.281	0.312	0.296	0.075	0.579	0.278	3.5
0.560	0.455	0.294	0.560	0.386	0.091	0.577	0.324	4
0.000	0.000	?	0.000	?	?	0.610	0.067	4.5
0.000	0.000	?	0.000	?	?	0.664	0.061	5

Weighted Avg. 0.257 0.204 ? 0.257 ? ? 0.584 0.221

=== Confusion Matrix ===

a	b	c	d	e	f	g	h	i	j	<-- classified as
0	0	7	0	0	15	3	68	0	0	| a = 0.5
0	0	25	0	0	37	43	106	0	0	| b = 1
0	0	73	0	0	26	29	83	0	0	| c = 1.5
0	0	116	2	0	133	144	245	0	0	| d = 2
0	0	101	0	0	172	253	328	0	0	| e = 2.5
0	0	176	3	0	332	582	806	0	0	| f = 3
0	0	153	3	0	331	642	928	0	0	| g = 3.5
0	0	197	0	0	328	481	1282	0	0	| h = 4
0	0	52	0	0	49	66	293	0	0	| i = 4.5
0	0	45	3	0	29	41	222	0	0	| j = 5

2. **NAIVEBAYES CLASSIFICATION:** Using the Bayes theorem, the Naive Bayes algorithm assumes that each characteristic has equal importance and that the categorized elements are independent of one another. As a result of its simple methodology that requires few training samples and treats all data points as significant, the Naive Bayes classifier is among the simplest choices.



Attribute	Class 0.5 (0.01)	1 (0.02)	1.5 (0.02)	2 (0.07)	2.5 (0.09)	3 (0.21)	3 (0.2
movieId							
mean	43114.0652	31728.5233	61683.2062	38383.146	32516.9983	27522.8282	24993.5
std. dev.	39858.0932	44052.5081	50371.1341	44683.755	41349.8466	38125.996	36432.4
weight sum	93	211	211	640	854	1899	2
precision	18.1118	18.1118	18.1118	18.1118	18.1118	18.1118	18.1
imdbId							
mean	551575.3086	574088.8521	1026716.054	632504.6909	514687.1795	440953.3894	390960.9
std. dev.	809009.0773	892862.4081	1022190.6847	816429.6621	737550.8812	677060.5405	638189.0
weight sum	93	211	211	640	854	1899	2
precision	640.1181	640.1181	640.1181	640.1181	640.1181	640.1181	640.1
tmdbId							
mean	51512.9535	52445.4154	82904.0486	50251.7076	39818.2427	35418.7157	29548.8
std. dev.	80670.7308	77190.3012	95896.655	69558.8814	62476.9864	56931.1663	52557.1
weight sum	93	211	211	640	854	1899	2
precision	46.0048	46.0048	46.0048	46.0048	46.0048	46.0048	46.0

Classifier output

```
Time taken to build model: 0 seconds

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

Correctly Classified Instances        2097               23.1636 %
Incorrectly Classified Instances      6956               76.8364 %
Kappa statistic                          0.0229
Mean absolute error                      0.162
Root mean squared error                  0.2944
Relative absolute error                 98.6023 %
Root relative squared error            102.7134 %
Total Number of Instances             9053

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.000    0.000    ?          0.000    ?          ?        0.584     0.018     0.5
                 0.000    0.000    ?          0.000    ?          ?        0.460     0.022     1
                 0.246    0.072    0.076      0.246    0.116      0.100    0.681     0.055     1.5
                 0.017    0.011    0.106      0.017    0.030      0.015    0.539     0.085     2
                 0.000    0.000    ?          0.000    ?          ?        0.499     0.096     2.5
                 0.000    0.000    ?          0.000    ?          ?        0.516     0.216     3
                 0.809    0.728    0.246      0.809    0.378      0.079    0.565     0.267     3.5
                 0.157    0.166    0.243      0.157    0.191      -0.010   0.520     0.265     4
                 0.000    0.000    ?          0.000    ?          ?        0.540     0.056     4.5
                 0.026    0.002    0.346      0.026    0.049      0.087    0.624     0.074     5
Weighted Avg.    0.232    0.210    ?          0.232    ?          ?        0.537     0.196
```

```
=== Confusion Matrix ===

    a    b    c    d    e    f    g    h    i    j   <-- classified as
    0    0    8    2    0    0   62   21    0    0 |    a = 0.5
    0    0   21    2    0    0  157   30    0    1 |    b = 1
    0    0   52    8    0    0   98   53    0    0 |    c = 1.5
    0    0   83   11    0    0  435  111    0    0 |    d = 2
    0    0   66    7    0    0  616  164    0    1 |    e = 2.5
    0    0  120   19    0    0 1464  291    0    5 |    f = 3
    0    0  105   14    0    0 1665  271    0    2 |    g = 3.5
    0    0  152   30    0    0 1740  360    0    6 |    h = 4
    0    0   38    7    0    0  329   84    0    2 |    i = 4.5
    0    0   41    4    0    0  190   96    0    9 |    j = 5
```

=== Run information ===

Scheme: weka.classifiers.bayes.NaiveBayes

Relation: original_data-weka.filters.unsupervised.attribute.NumericToNominal-R4

Instances: 9053

Attributes: 4

movieId

imdbId

tmdbId

average_rating

Test mode: 10-fold cross-validation

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

Naive Bayes Classifier

Class

Attribute 0.5 1 1.5 2 2.5 3 3.5 4 4.5 5

(0.01) (0.02) (0.02) (0.07) (0.09) (0.21) (0.23) (0.25) (0.05) (0.04)

==========

movieId

mean 43114.0652 31728.5233 61683.2062 38383.146 32516.9983 27522.8282 24993.5775 28747.4614 34546.0901 50022.8165

std. dev. 39858.0932 44052.5081 50371.1341 44683.755 41349.8466 38125.996 36432.4383 38352.7922 39779.818 50896.4714

weight sum 93 211 211 640 854 1899 2057 2288 460 340

precision 18.1118 18.1118 18.1118 18.1118 18.1118 18.1118 18.1118 18.1118 18.1118 18.1118

imdbId

mean 551575.3086 574088.8521 1026716.054 632504.6909 514687.1795 440953.3894 390960.9179 438107.6797 478897.3953 570542.9066

std. dev. 809009.0773 892862.4081 1022190.6847 816429.6621 737550.8812 677060.5405 638189.0275 702674.1148 798591.5249 903914.6618

weight sum 93 211 211 640 854 1899 2057 2288 460 340

precision 640.1181 640.1181 640.1181 640.1181 640.1181 640.1181 640.1181 640.1181 640.1181 640.1181

tmdbId

mean 51512.9535 52445.4154 82904.0486 50251.7076 39818.2427 35418.7157 29548.8936 36265.1144 40992.0327 54627.9349

std. dev. 80670.7308 77190.3012 95896.655 69558.8814 62476.9864 56931.1663 52557.1259 58783.3615 61226.9475 71746.0677

weight sum 93 211 211 640 854 1899 2057 2288 460 340

precision 46.0048 46.0048 46.0048 46.0048 46.0048 46.0048 46.0048 46.0048 46.0048 46.0048

Time taken to build model: 0.03 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	2097	23.1636 %
Incorrectly Classified Instances	6956	76.8364 %
Kappa statistic	0.0229	
Mean absolute error	0.162	
Root mean squared error	0.2944	
Relative absolute error	98.6023 %	
Root relative squared error	102.7134 %	
Total Number of Instances	9053	

=== Detailed Accuracy By Class ===

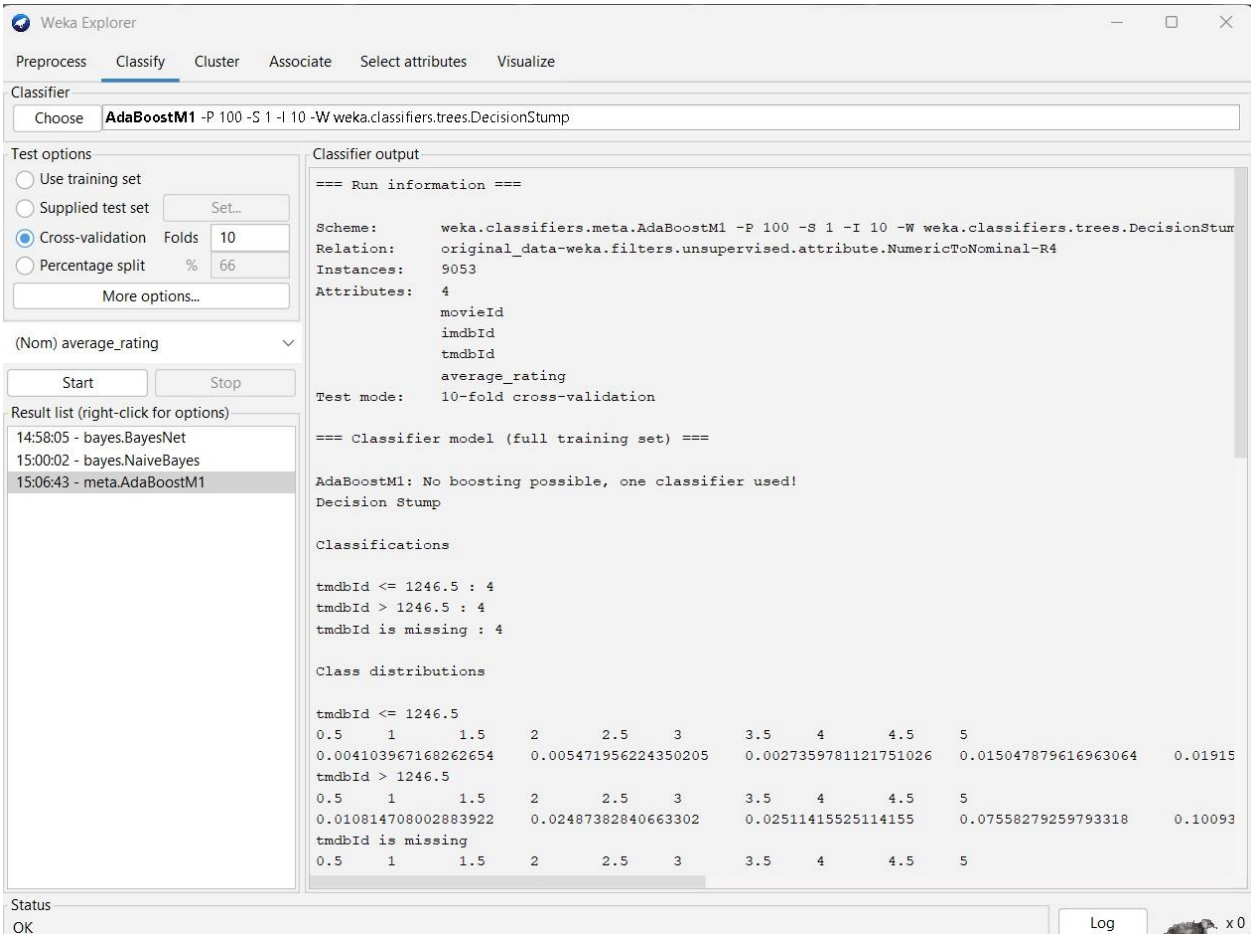
TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0.000	0.000	?	0.000	?	?	0.584	0.018	0.5
0.000	0.000	?	0.000	?	?	0.460	0.022	1
0.246	0.072	0.076	0.246	0.116	0.100	0.681	0.055	1.5
0.017	0.011	0.106	0.017	0.030	0.015	0.539	0.085	2
0.000	0.000	?	0.000	?	?	0.499	0.096	2.5
0.000	0.000	?	0.000	?	?	0.516	0.216	3

	0.809	0.728	0.246	0.809	0.378	0.079	0.565	0.267	3.5
	0.157	0.166	0.243	0.157	0.191	-0.010	0.520	0.265	4
	0.000	0.000	?	0.000	?	?	0.540	0.056	4.5
	0.026	0.002	0.346	0.026	0.049	0.087	0.624	0.074	5
Weighted Avg.	0.232	0.210	?	0.232	?	?	0.537	0.196	

=== Confusion Matrix ===

a	b	c	d	e	f	g	h	i	j	<-- classified as
0	0	8	2	0	0	62	21	0	0	a = 0.5
0	0	21	2	0	0	157	30	0	1	b = 1
0	0	52	8	0	0	98	53	0	0	c = 1.5
0	0	83	11	0	0	435	111	0	0	d = 2
0	0	66	7	0	0	616	164	0	1	e = 2.5
0	0	120	19	0	0	1464	291	0	5	f = 3
0	0	105	14	0	0	1665	271	0	2	g = 3.5
0	0	152	30	0	0	1740	360	0	6	h = 4
0	0	38	7	0	0	329	84	0	2	i = 4.5
0	0	41	4	0	0	190	96	0	9	j = 5

3. **ADABOOST M1 CLASSIFICATION:** AdaBoost, is a technique used to increase decision trees' efficacy. It works well as a robust classifier when combined with less proficient learners. AdaBoost improves classification accuracy without requiring significant parameter changes, making it a widely used and adaptable classification technique.



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

Correctly Classified Instances        2292               25.3176 %
Incorrectly Classified Instances      6761               74.6824 %
Kappa statistic                          0.0019
Mean absolute error                      0.1635
Root mean squared error                  0.286
Relative absolute error                 99.5181 %
Root relative squared error             99.8045 %
Total Number of Instances             9053

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0.000    0.000    ?          0.000    ?          ?        0.569     0.012     0.5
               0.000    0.000    ?          0.000    ?          ?        0.533     0.025     1
               0.000    0.000    ?          0.000    ?          ?        0.563     0.027     1.5
               0.000    0.000    ?          0.000    ?          ?        0.529     0.074     2
               0.000    0.000    ?          0.000    ?          ?        0.525     0.099     2.5
               0.000    0.000    ?          0.000    ?          ?        0.516     0.215     3
               0.047    0.037    0.272      0.047    0.080      0.022    0.521     0.242     3.5
               0.959    0.961    0.252      0.959    0.400      -0.004   0.540     0.291     4
               0.000    0.000    ?          0.000    ?          ?        0.526     0.056     4.5
               0.000    0.000    ?          0.000    ?          ?        0.546     0.042     5
Weighted Avg.  0.253    0.251    ?          0.253    ?          ?        0.529     0.194
```

```
=== Confusion Matrix ===

    a    b    c    d    e    f    g    h    i    j   <-- classified as
    0    0    0    0    0    0    3   90    0    0 |    a = 0.5
    0    0    0    0    0    0    6  205    0    0 |    b = 1
    0    0    0    0    0    0    4  207    0    0 |    c = 1.5
    0    0    0    0    0    0   27  613    0    0 |    d = 2
    0    0    0    0    0    0   32  822    0    0 |    e = 2.5
    0    0    0    0    0    0   74 1825    0    0 |    f = 3
    0    0    0    0    0    0   97 1960    0    0 |    g = 3.5
    0    0    0    0    0    0   93 2195    0    0 |    h = 4
    0    0    0    0    0    0   14  446    0    0 |    i = 4.5
    0    0    0    0    0    0    7  333    0    0 |    j = 5
```

=== Run information ===

Scheme: weka.classifiers.meta.AdaBoostM1 -P 100 -S 1 -I 10 -W weka.classifiers.trees.DecisionStump

Relation: original_data-weka.filters.unsupervised.attribute.NumericToNominal-R4

Instances: 9053

Attributes: 4

movieId

imdbId

tmdbId

average_rating

Test mode: 10-fold cross-validation

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

AdaBoostM1: No boosting possible, one classifier used!

Decision Stump

Classifications

tmdbId <= 1246.5 : 4

tmdbId > 1246.5 : 4

tmdbId is missing : 4

Class distributions

tmdbId <= 1246.5

0.5 1 1.5 2 2.5 3 3.5 4 4.5 5

0.004103967168262654 0.005471956224350205 0.0027359781121751026 0.015047879616963064 0.019151846785225718 0.1409028727770178 0.3023255813953488 0.43091655266757867 0.06839945280437756 0.01094391244870041

tmdbId > 1246.5

0.5 1 1.5 2 2.5 3 3.5 4 4.5 5

0.010814708002883922 0.02487382840663302 0.02511415525114155
0.07558279259793318 0.10093727469358327 0.21581350636866137
0.220620043258832 0.2370824321076664 0.049267003124248976
0.039894256188416244

tmdbId is missing

0.5 1 1.5 2 2.5 3 3.5 4 4.5 5

0.010272837733348061 0.023307190986413345 0.023307190986413345
0.07069479730476086 0.09433337015354026 0.20976471887772008
0.22721749696233293 0.25273390036452004 0.050811885562796866
0.0375566110681542

Time taken to build model: 0.01 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	2292	25.3176 %
Incorrectly Classified Instances	6761	74.6824 %
Kappa statistic	0.0019	
Mean absolute error	0.1635	
Root mean squared error	0.286	
Relative absolute error	99.5181 %	
Root relative squared error	99.8045 %	
Total Number of Instances	9053	

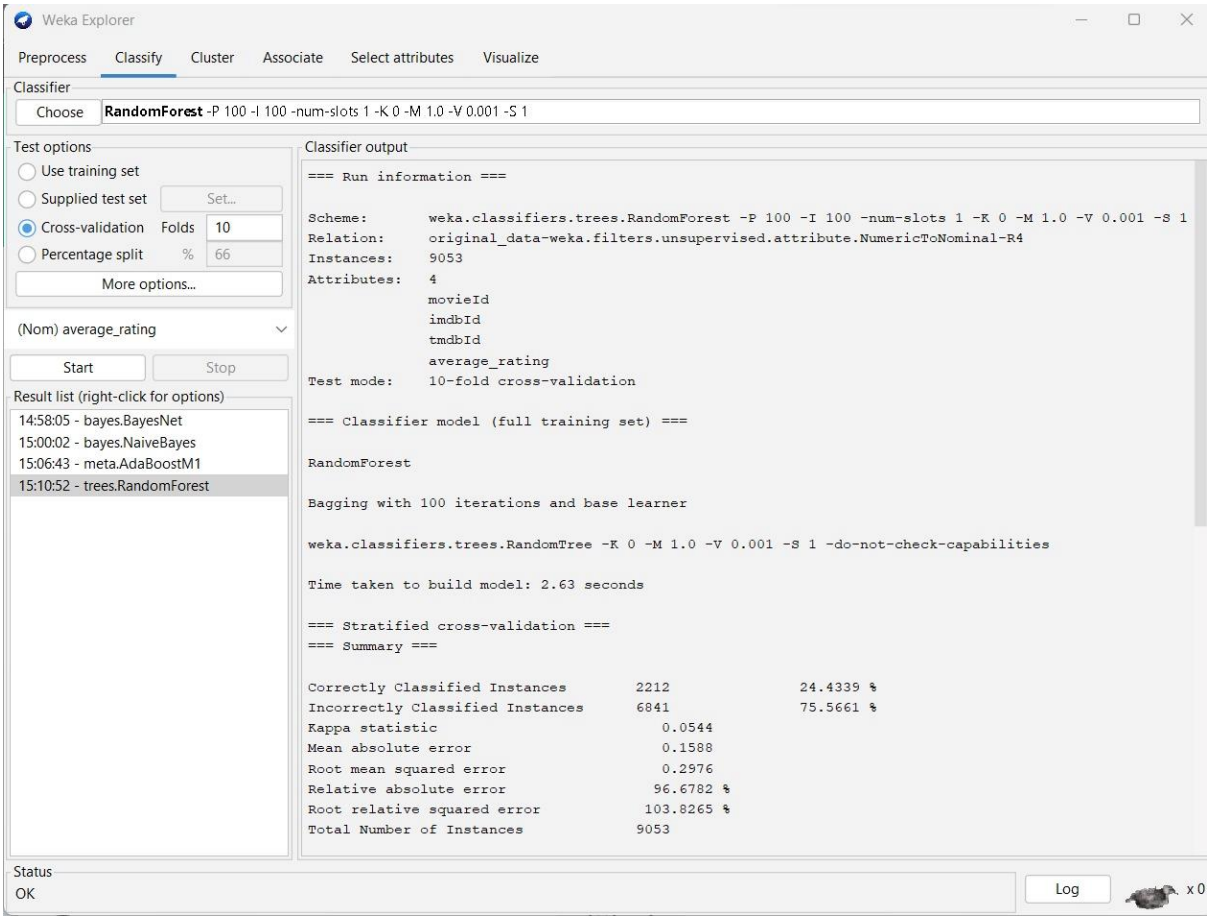
=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0.000	0.000	?	0.000	?	?	0.569	0.012	0.5
0.000	0.000	?	0.000	?	?	0.533	0.025	1
0.000	0.000	?	0.000	?	?	0.563	0.027	1.5
0.000	0.000	?	0.000	?	?	0.529	0.074	2
0.000	0.000	?	0.000	?	?	0.525	0.099	2.5
0.000	0.000	?	0.000	?	?	0.516	0.215	3
0.047	0.037	0.272	0.047	0.080	0.022	0.521	0.242	3.5
0.959	0.961	0.252	0.959	0.400	-0.004	0.540	0.291	4
0.000	0.000	?	0.000	?	?	0.526	0.056	4.5
0.000	0.000	?	0.000	?	?	0.546	0.042	5
Weighted Avg. 0.253	0.251	?	0.253	?	?	0.529	0.194	

=== Confusion Matrix ===

a	b	c	d	e	f	g	h	i	j	<-- classified as
0	0	0	0	0	0	3	90	0	0	| a = 0.5
0	0	0	0	0	0	6	205	0	0	| b = 1
0	0	0	0	0	0	4	207	0	0	| c = 1.5
0	0	0	0	0	0	27	613	0	0	| d = 2
0	0	0	0	0	0	32	822	0	0	| e = 2.5
0	0	0	0	0	0	74	1825	0	0	| f = 3
0	0	0	0	0	0	97	1960	0	0	| g = 3.5
0	0	0	0	0	0	93	2195	0	0	| h = 4
0	0	0	0	0	0	14	446	0	0	| i = 4.5
0	0	0	0	0	0	7	333	0	0	| j = 5

4. **RANDOM FOREST CLASSIFICATION:** Regression detection and prediction are two areas in which Random Forest is a flexible approach. The most ideal decision tree is determined via a voting system that generates many decision trees, much like a forest. Random Forest is not like other classifiers since it can handle overfitting well, which makes its predictions more accurate.



```
=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.032    0.003    0.100      0.032    0.049      0.051    0.624     0.021     0.5
                 0.014    0.010    0.033      0.014    0.020      0.006    0.594     0.029     1
                 0.024    0.011    0.049      0.024    0.032      0.018    0.605     0.035     1.5
                 0.075    0.046    0.110      0.075    0.089      0.034    0.577     0.090     2
                 0.093    0.068    0.124      0.093    0.106      0.028    0.561     0.114     2.5
                 0.286    0.224    0.254      0.286    0.269      0.060    0.556     0.241     3
                 0.317    0.250    0.272      0.317    0.293      0.064    0.564     0.263     3.5
                 0.361    0.292    0.295      0.361    0.325      0.065    0.570     0.329     4
                 0.028    0.023    0.062      0.028    0.039      0.008    0.591     0.068     4.5
                 0.109    0.018    0.191      0.109    0.139      0.119    0.668     0.124     5
Weighted Avg.    0.244    0.190    0.222      0.244    0.230      0.055    0.572     0.220

=== Confusion Matrix ===

   a   b   c   d   e   f   g   h   i   j   <-- classified as
   3   0   1   6   4  19  22  35   1   2 |   a = 0.5
   0   3   6  14  19  60  44  60   0   5 |   b = 1
   2   9   5  21  18  49  40  63   1   3 |   c = 1.5
   2   8  14  48  57 182 139 163  12  15 |   d = 2
   1   8   6  60  79 235 232 199  24  10 |   e = 2.5
   1  23  25  98 151 544 490 503  30  34 |   f = 3
  10  18  18  75 149 447 653 617  49  21 |   g = 3.5
   7  15  25  82 130 469 615 827  66  52 |   h = 4
   1   3   0  15  17  77 111 208  13  15 |   i = 4.5
   3   4   2  19  15  63  55 129  13  37 |   j = 5
```

=== Run information ===

Scheme: weka.classifiers.trees.RandomForest -P 100 -I 100 -num-slots 1 -K 0 -M 1.0 -V 0.001 -S 1

Relation: original_data-weka.filters.unsupervised.attribute.NumericToNominal-R4

Instances: 9053

Attributes: 4

movieId

imdbId

tmdbId

average_rating

Test mode: 10-fold cross-validation

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

RandomForest

Bagging with 100 iterations and base learner

weka.classifiers.trees.RandomTree -K 0 -M 1.0 -V 0.001 -S 1 -do-not-check-capabilities

Time taken to build model: 2.46 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	2212	24.4339 %
Incorrectly Classified Instances	6841	75.5661 %
Kappa statistic	0.0544	
Mean absolute error	0.1588	
Root mean squared error	0.2976	
Relative absolute error	96.6782 %	
Root relative squared error	103.8265 %	
Total Number of Instances	9053	

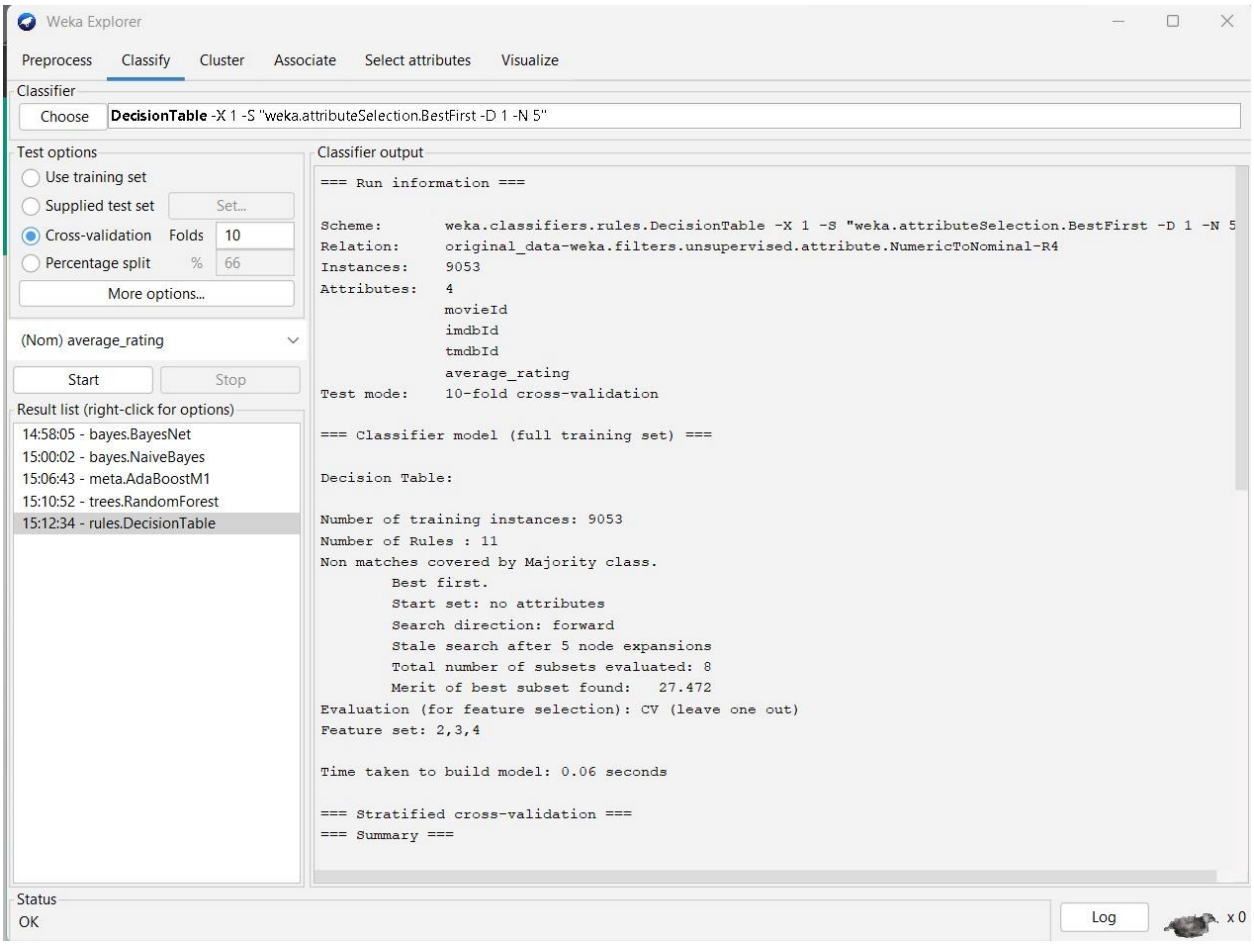
=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.032	0.003	0.100	0.032	0.049	0.051	0.624	0.021	0.5
	0.014	0.010	0.033	0.014	0.020	0.006	0.594	0.029	1
	0.024	0.011	0.049	0.024	0.032	0.018	0.605	0.035	1.5
	0.075	0.046	0.110	0.075	0.089	0.034	0.577	0.090	2
	0.093	0.068	0.124	0.093	0.106	0.028	0.561	0.114	2.5
	0.286	0.224	0.254	0.286	0.269	0.060	0.556	0.241	3
	0.317	0.250	0.272	0.317	0.293	0.064	0.564	0.263	3.5
	0.361	0.292	0.295	0.361	0.325	0.065	0.570	0.329	4
	0.028	0.023	0.062	0.028	0.039	0.008	0.591	0.068	4.5
	0.109	0.018	0.191	0.109	0.139	0.119	0.668	0.124	5
Weighted Avg.	0.244	0.190	0.222	0.244	0.230	0.055	0.572	0.220	

=== Confusion Matrix ===

a	b	c	d	e	f	g	h	i	j	<-- classified as
3	0	1	6	4	19	22	35	1	2	a = 0.5
0	3	6	14	19	60	44	60	0	5	b = 1
2	9	5	21	18	49	40	63	1	3	c = 1.5
2	8	14	48	57	182	139	163	12	15	d = 2
1	8	6	60	79	235	232	199	24	10	e = 2.5
1	23	25	98	151	544	490	503	30	34	f = 3
10	18	18	75	149	447	653	617	49	21	g = 3.5
7	15	25	82	130	469	615	827	66	52	h = 4
1	3	0	15	17	77	111	208	13	15	i = 4.5
3	4	2	19	15	63	55	129	13	37	j = 5

5. **DECISION TREE CLASSIFICATION:** Building a tree-shaped decision-making model by repeated dataset division into subgroups based on criteria such as information gain or Gini index is known as decision tree classification. This methodology provides a clear and comprehensible means of categorizing data into groups.



```
Time taken to build model: 0.06 seconds

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

Correctly Classified Instances        2469               27.2727 %
Incorrectly Classified Instances      6584               72.7273 %
Kappa statistic                          0.0491
Mean absolute error                      0.1621
Root mean squared error                  0.2847
Relative absolute error                 98.6324 %
Root relative squared error             99.3325 %
Total Number of Instances             9053

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.000    0.000    ?          0.000    ?          ?        0.685     0.017     0.5
                 0.000    0.000    ?          0.000    ?          ?        0.608     0.032     1
                 0.000    0.000    ?          0.000    ?          ?        0.682     0.048     1.5
                 0.000    0.000    0.000      0.000    0.000      -0.004   0.596     0.091     2
                 0.000    0.002    0.000      0.000    0.000      -0.012   0.566     0.106     2.5
                 0.291    0.243    0.241      0.291    0.264      0.045    0.558     0.238     3
                 0.243    0.171    0.294      0.243    0.266      0.077    0.580     0.279     3.5
                 0.611    0.529    0.281      0.611    0.384      0.071    0.580     0.328     4
                 0.000    0.000    ?          0.000    ?          ?        0.636     0.084     4.5
                 0.059    0.006    0.290      0.059    0.098      0.116    0.676     0.091     5
Weighted Avg.    0.273    0.224    ?          0.273    ?          ?        0.586     0.222
```

```
=== Confusion Matrix ===

    a    b    c    d    e    f    g    h    i    j   <-- classified as
    0    0    0    0    0   18    4   71    0    0 |    a = 0.5
    0    0    0    0    0   60   26  124    0    1 |    b = 1
    0    0    0    0    1   41   16  151    0    2 |    c = 1.5
    0    0    0    0    1  182  112  342    0    3 |    d = 2
    0    0    0    0    0  255  193  401    0    5 |    e = 2.5
    0    0    0    0    3  553  425  908    0   10 |    f = 3
    0    0    0    2    4  527  499 1013    0   12 |    g = 3.5
    0    0    0    0    2  522  356 1397    0   11 |    h = 4
    0    0    0    0    0   75   45  335    0    5 |    i = 4.5
    0    0    0    0    2   60   21  237    0   20 |    j = 5
```

=== Run information ===

Scheme: weka.classifiers.rules.DecisionTable -X 1 -S "weka.attributeSelection.BestFirst -D 1 -N 5"

Relation: original_data-weka.filters.unsupervised.attribute.NumericToNominal-R4

Instances: 9053

Attributes: 4

movieId

imdbId

tmdbId

average_rating

Test mode: 10-fold cross-validation

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

Decision Table:

Number of training instances: 9053

Number of Rules : 11

Non matches covered by Majority class.

Best first.

Start set: no attributes

Search direction: forward

Stale search after 5 node expansions

Total number of subsets evaluated: 8

Merit of best subset found: 27.472

Evaluation (for feature selection): CV (leave one out)

Feature set: 2,3,4

Time taken to build model: 0.14 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	2469	27.2727 %
Incorrectly Classified Instances	6584	72.7273 %
Kappa statistic	0.0491	
Mean absolute error	0.1621	
Root mean squared error	0.2847	
Relative absolute error	98.6324 %	
Root relative squared error	99.3325 %	
Total Number of Instances	9053	

=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0.000	0.000	?	0.000	?	?	0.685	0.017	0.5
0.000	0.000	?	0.000	?	?	0.608	0.032	1
0.000	0.000	?	0.000	?	?	0.682	0.048	1.5
0.000	0.000	0.000	0.000	0.000	-0.004	0.596	0.091	2
0.000	0.002	0.000	0.000	0.000	-0.012	0.566	0.106	2.5
0.291	0.243	0.241	0.291	0.264	0.045	0.558	0.238	3
0.243	0.171	0.294	0.243	0.266	0.077	0.580	0.279	3.5
0.611	0.529	0.281	0.611	0.384	0.071	0.580	0.328	4
0.000	0.000	?	0.000	?	?	0.636	0.084	4.5
0.059	0.006	0.290	0.059	0.098	0.116	0.676	0.091	5
Weighted Avg. 0.273	0.224	?	0.273	?	?	0.586	0.222	

=== Confusion Matrix ===

a	b	c	d	e	f	g	h	i	j	<-- classified as
0	0	0	0	0	18	4	71	0	0	a = 0.5
0	0	0	0	0	60	26	124	0	1	b = 1
0	0	0	0	1	41	16	151	0	2	c = 1.5
0	0	0	0	1	182	112	342	0	3	d = 2
0	0	0	0	0	255	193	401	0	5	e = 2.5
0	0	0	0	3	553	425	908	0	10	f = 3
0	0	0	2	4	527	499	1013	0	12	g = 3.5
0	0	0	0	2	522	356	1397	0	11	h = 4
0	0	0	0	0	75	45	335	0	5	i = 4.5
0	0	0	0	2	60	21	237	0	20	j = 5

Classifiers Using WEKA On Extended Dataset

1. **BAYESNET CLASSIFICATION:**

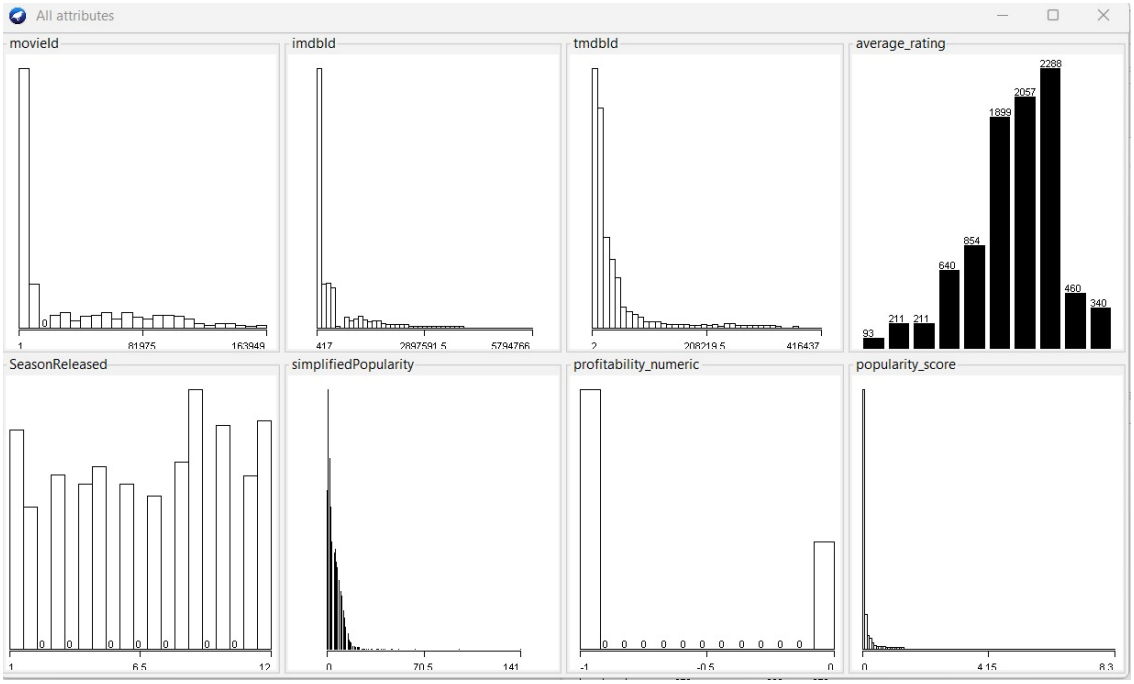


Fig: Dataset's visual representation following data transformation

Upon clicking the Edit button, we can observe the alterations in the data following the preprocessing steps.

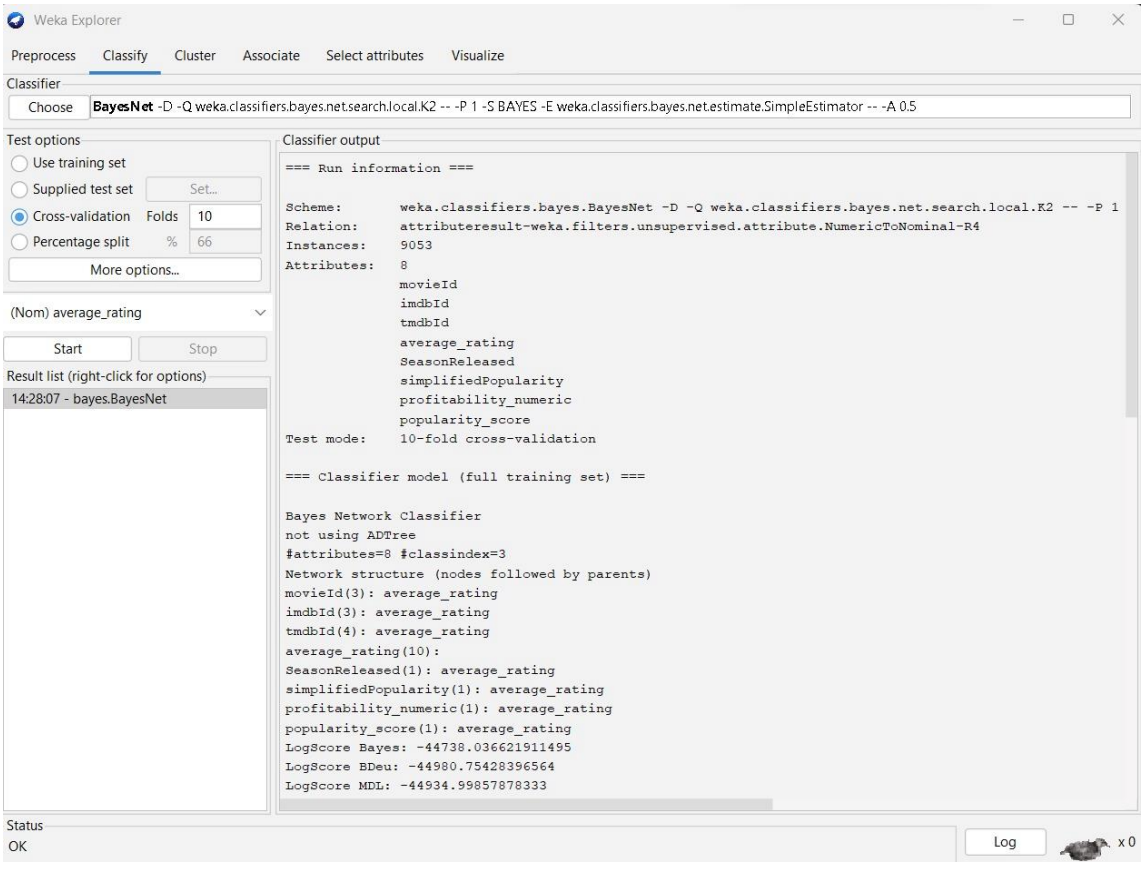
Viewer

Relation: Extended_Data-weka.filters.unsupervised.attribute.NumericToNominal-R4

No.	1: movieId Numeric	2: imdbId Numeric	3: tmdbId Numeric	4: average_rating Nominal	5: SeasonReleased Numeric	6: simplifiedPopularity Numeric	7: profitability_numeric Numeric	8: **popularity_score** Numeric
1	1.0	114709.0	862.0	4	10.0	22.0	0.0	4.31
2	2.0	113497.0	8844.0	3.5	12.0	17.0	0.0	1.72
3	3.0	113228.0	15602.0	3	12.0	12.0	-1.0	0.06
4	4.0	114885.0	31357.0	2.5	12.0	4.0	0.0	0.02
5	5.0	113041.0	11862.0	3.5	2.0	8.0	0.0	0.1
6	6.0	113277.0	949.0	4	12.0	18.0	0.0	1.5
7	7.0	114319.0	11860.0	3.5	12.0	7.0	-1.0	0.09
8	8.0	112302.0	45325.0	4	12.0	3.0	-1.0	0.03
9	9.0	114576.0	9091.0	3	12.0	5.0	0.0	0.1
10	10.0	113189.0	710.0	3.5	11.0	15.0	0.0	0.81
11	11.0	112346.0	9087.0	3.5	11.0	6.0	0.0	0.13
12	12.0	112896.0	12110.0	3	12.0	5.0	-1.0	0.12
13	13.0	112453.0	21032.0	4	12.0	12.0	0.0	0.31
14	14.0	113987.0	10858.0	3.5	12.0	5.0	-1.0	0.05
15	15.0	112760.0	1408.0	2.5	12.0	7.0	-1.0	0.08
16	16.0	112641.0	524.0	4	11.0	10.0	0.0	1.08
17	17.0	114388.0	4584.0	4	12.0	11.0	0.0	0.27
18	18.0	113101.0	5.0	3.5	12.0	9.0	0.0	0.36
19	19.0	112281.0	9273.0	2.5	11.0	8.0	0.0	0.71
20	20.0	113845.0	11517.0	2.5	11.0	7.0	-1.0	0.12
21	21.0	113161.0	8012.0	3.5	10.0	13.0	0.0	0.2
22	22.0	112722.0	1710.0	3.5	10.0	11.0	-1.0	0.13
23	23.0	112401.0	9691.0	3	10.0	11.0	-1.0	0.24
24	24.0	114168.0	12665.0	3	10.0	12.0	-1.0	0.09

Add instance | Undo | OK | Cancel

Fig: Data after preprocessing



```
Time taken to build model: 0.04 seconds

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

Correctly Classified Instances        2331               25.7484 %
Incorrectly Classified Instances      6722               74.2516 %
Kappa statistic                          0.0538
Mean absolute error                      0.1619
Root mean squared error                  0.2866
Relative absolute error                 98.5507 %
Root relative squared error            100.008  %
Total Number of Instances             9053

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.000    0.000    ?          0.000    ?          ?        0.700     0.018     0.5
                 0.000    0.000    ?          0.000    ?          ?        0.614     0.032     1
                 0.346    0.099    0.077      0.346    0.126      0.122    0.706     0.052     1.5
                 0.003    0.001    0.182      0.003    0.006      0.015    0.599     0.093     2
                 0.000    0.000    ?          0.000    ?          ?        0.561     0.107     2.5
                 0.175    0.157    0.229      0.175    0.198      0.020    0.559     0.244     3
                 0.312    0.235    0.281      0.312    0.296      0.075    0.579     0.278     3.5
                 0.560    0.455    0.294      0.560    0.386      0.091    0.577     0.324     4
                 0.000    0.000    ?          0.000    ?          ?        0.610     0.067     4.5
                 0.000    0.000    ?          0.000    ?          ?        0.664     0.061     5
Weighted Avg.    0.257    0.204    ?          0.257    ?          ?        0.584     0.221
```

Classifier output

```
Root relative squared error            100.008  %
Total Number of Instances             9053

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.000    0.000    ?          0.000    ?          ?        0.700     0.018     0.5
                 0.000    0.000    ?          0.000    ?          ?        0.614     0.032     1
                 0.346    0.099    0.077      0.346    0.126      0.122    0.706     0.052     1.5
                 0.003    0.001    0.182      0.003    0.006      0.015    0.599     0.093     2
                 0.000    0.000    ?          0.000    ?          ?        0.561     0.107     2.5
                 0.175    0.157    0.229      0.175    0.198      0.020    0.559     0.244     3
                 0.312    0.235    0.281      0.312    0.296      0.075    0.579     0.278     3.5
                 0.560    0.455    0.294      0.560    0.386      0.091    0.577     0.324     4
                 0.000    0.000    ?          0.000    ?          ?        0.610     0.067     4.5
                 0.000    0.000    ?          0.000    ?          ?        0.664     0.061     5
Weighted Avg.    0.257    0.204    ?          0.257    ?          ?        0.584     0.221

=== Confusion Matrix ===

    a    b    c    d    e    f    g    h    i    j   <-- classified as
    0    0    7    0    0   15    3   68    0    0 |    a = 0.5
    0    0   25    0    0   37   43  106    0    0 |    b = 1
    0    0   73    0    0   26   29   83    0    0 |    c = 1.5
    0    0  116    2    0  133  144  245    0    0 |    d = 2
    0    0  101    0    0  172  253  328    0    0 |    e = 2.5
    0    0  176    3    0  332  582  806    0    0 |    f = 3
    0    0  153    3    0  331  642  928    0    0 |    g = 3.5
    0    0  197    0    0  328  481 1282    0    0 |    h = 4
    0    0   52    0    0   49   66  293    0    0 |    i = 4.5
    0    0   45    3    0   29   41  222    0    0 |    j = 5
```

=== Run information ===

Scheme: weka.classifiers.bayes.BayesNet -D -Q weka.classifiers.bayes.net.search.local.K2 --
-P 1 -S BAYES -E weka.classifiers.bayes.net.estimate.SimpleEstimator -- -A 0.5

Relation: attributeresult-weka.filters.unsupervised.attribute.NumericToNominal-R4

Instances: 9053

Attributes: 8

movieId

imdbId

tmdbId

average_rating

SeasonReleased

simplifiedPopularity

profitability_numeric

popularity_score

Test mode: 10-fold cross-validation

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

Bayes Network Classifier

not using ADTree

#attributes=8 #classindex=3

Network structure (nodes followed by parents)

movieId(3): average_rating

imdbId(3): average_rating

tmdbId(4): average_rating

average_rating(10):

SeasonReleased(1): average_rating

simplifiedPopularity(1): average_rating

profitability_numeric(1): average_rating

popularity_score(1): average_rating

LogScore Bayes: -44738.036621911495

LogScore BDeu: -44980.75428396564

LogScore MDL: -44934.99857878333

LogScore ENTROPY: -44575.11994558104

LogScore AIC: -44654.11994558104

Time taken to build model: 0.04 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	2331	25.7484 %
Incorrectly Classified Instances	6722	74.2516 %
Kappa statistic	0.0538	
Mean absolute error	0.1619	
Root mean squared error	0.2866	
Relative absolute error	98.5507 %	
Root relative squared error	100.008 %	
Total Number of Instances	9053	

=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0.000	0.000	?	0.000	?	?	0.700	0.018	0.5
0.000	0.000	?	0.000	?	?	0.614	0.032	1
0.346	0.099	0.077	0.346	0.126	0.122	0.706	0.052	1.5
0.003	0.001	0.182	0.003	0.006	0.015	0.599	0.093	2
0.000	0.000	?	0.000	?	?	0.561	0.107	2.5
0.175	0.157	0.229	0.175	0.198	0.020	0.559	0.244	3
0.312	0.235	0.281	0.312	0.296	0.075	0.579	0.278	3.5
0.560	0.455	0.294	0.560	0.386	0.091	0.577	0.324	4
0.000	0.000	?	0.000	?	?	0.610	0.067	4.5
0.000	0.000	?	0.000	?	?	0.664	0.061	5
Weighted Avg. 0.257	0.204	?	0.257	?	?	0.584	0.221	

=== Confusion Matrix ===

a	b	c	d	e	f	g	h	i	j	<-- classified as
0	0	7	0	0	15	3	68	0	0	| a = 0.5
0	0	25	0	0	37	43	106	0	0	| b = 1
0	0	73	0	0	26	29	83	0	0	| c = 1.5
0	0	116	2	0	133	144	245	0	0	| d = 2
0	0	101	0	0	172	253	328	0	0	| e = 2.5
0	0	176	3	0	332	582	806	0	0	| f = 3
0	0	153	3	0	331	642	928	0	0	| g = 3.5
0	0	197	0	0	328	481	1282	0	0	| h = 4
0	0	52	0	0	49	66	293	0	0	| i = 4.5
0	0	45	3	0	29	41	222	0	0	| j = 5

2. NAIVE BAYES CLASSIFICATION:

Classifier output

```
=== Run information ===

Scheme:       weka.classifiers.bayes.NaiveBayes
Relation:     attributeresult-weka.filters.unsupervised.attribute.NumericToNominal-R4
Instances:    9053
Attributes:   8
              movieId
              imdbId
              tmdbId
              average_rating
              SeasonReleased
              simplifiedPopularity
              profitability_numeric
              popularity_score
Test mode:    10-fold cross-validation

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

Naive Bayes Classifier

                            Class
Attribute                     0.5            1          1.5            2          2.5            3
                            (0.01)       (0.02)       (0.02)       (0.07)       (0.09)       (0.21)
=============================================================================================================
movieId
  mean                  43114.0652   31728.5233   61683.2062    38383.146   32516.9983   27522.8282
  std. dev.             39858.0932   44052.5081   50371.1341    44683.755   41349.8466    38125.996
  weight sum                    93          211          211          640          854         1899
  precision                18.1118      18.1118      18.1118      18.1118      18.1118      18.1118

imdbId
  mean                  551575.3086  574088.8521  1026716.054  632504.6909  514687.1795  440953.3894
  std. dev.             809009.0773  892862.4081 1022190.6847  816429.6621  737550.8812  677060.5405
```

Classifier output

tmdbId						
mean	51512.9535	52445.4154	82904.0486	50251.7076	39818.2427	35418.7157
std. dev.	80670.7308	77190.3012	95896.655	69558.8814	62476.9864	56931.1663
weight sum	93	211	211	640	854	1899
precision	46.0048	46.0048	46.0048	46.0048	46.0048	46.0048
SeasonReleased						
mean	6.4946	6.6114	6.4408	6.9515	6.6741	6.7155
std. dev.	3.597	3.5346	3.6804	3.4653	3.5552	3.5059
weight sum	93	211	211	639	853	1898
precision	1	1	1	1	1	1
simplifiedPopularity						
mean	4.6862	4.9511	3.8728	4.9971	4.9382	5.0692
std. dev.	3.8626	4.4579	3.9391	4.4384	4.7013	5.9958
weight sum	93	211	211	640	854	1899
precision	3.2045	3.2045	3.2045	3.2045	3.2045	3.2045
profitability_numeric						
mean	-0.7419	-0.7156	-0.7156	-0.7375	-0.692	-0.7004
std. dev.	0.4376	0.4511	0.4511	0.44	0.4617	0.4581
weight sum	93	211	211	640	854	1899
precision	1	1	1	1	1	1
popularity_score						
mean	0.0927	0.1224	0.0644	0.1281	0.1221	0.1339
std. dev.	0.1722	0.3835	0.1487	0.3851	0.4465	0.4321
weight sum	93	211	211	640	854	1899
precision	0.0323	0.0323	0.0323	0.0323	0.0323	0.0323

Classifier output

```
Time taken to build model: 0.01 seconds

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

Correctly Classified Instances        2049               22.6334 %
Incorrectly Classified Instances      7004               77.3666 %
Kappa statistic                          0.0266
Mean absolute error                      0.1641
Root mean squared error                  0.2974
Relative absolute error                 99.877  %
Root relative squared error            103.7542 %
Total Number of Instances             9053

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.000    0.000    ?          0.000    ?          ?        0.550     0.013     0.5
                 0.000    0.000    ?          0.000    ?          ?        0.446     0.021     1
                 0.351    0.107    0.072      0.351    0.120      0.116    0.674     0.054     1.5
                 0.006    0.002    0.160      0.006    0.012      0.018    0.519     0.081     2
                 0.000    0.000    ?          0.000    ?          ?        0.505     0.096     2.5
                 0.085    0.081    0.218      0.085    0.123      0.006    0.510     0.214     3
                 0.728    0.645    0.249      0.728    0.371      0.073    0.557     0.256     3.5
                 0.133    0.135    0.250      0.133    0.174      -0.003   0.511     0.261     4
                 0.000    0.000    ?          0.000    ?          ?        0.531     0.056     4.5
                 0.021    0.003    0.241      0.021    0.038      0.061    0.592     0.080     5
Weighted Avg.    0.226    0.201    ?          0.226    ?          ?        0.528     0.192
```

Classifier output

```
Root relative squared error            103.7542 %
Total Number of Instances             9053

=== Detailed Accuracy By Class ===

                TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                0.000    0.000    ?          0.000    ?          ?        0.550     0.013     0.5
                0.000    0.000    ?          0.000    ?          ?        0.446     0.021     1
                0.351    0.107    0.072      0.351    0.120      0.116    0.674     0.054     1.5
                0.006    0.002    0.160      0.006    0.012      0.018    0.519     0.081     2
                0.000    0.000    ?          0.000    ?          ?        0.505     0.096     2.5
                0.085    0.081    0.218      0.085    0.123      0.006    0.510     0.214     3
                0.728    0.645    0.249      0.728    0.371      0.073    0.557     0.256     3.5
                0.133    0.135    0.250      0.133    0.174      -0.003   0.511     0.261     4
                0.000    0.000    ?          0.000    ?          ?        0.531     0.056     4.5
                0.021    0.003    0.241      0.021    0.038      0.061    0.592     0.080     5
Weighted Avg.   0.226    0.201    ?          0.226    ?          ?        0.528     0.192

=== Confusion Matrix ===

    a    b    c    d    e    f    g    h    i    j   <-- classified as
    0    0   11    0    0    4   54   24    0    0 |    a = 0.5
    0    0   29    0    0   19  138   24    0    1 |    b = 1
    0    0   74    1    0   14   85   37    0    0 |    c = 1.5
    0    0  117    4    0   50  384   83    0    2 |    d = 2
    0    0  105    2    0   71  554  120    0    2 |    e = 2.5
    0    0  184    3    0  162 1299  248    0    3 |    f = 3
    0    0  154    2    0  171 1497  227    0    6 |    g = 3.5
    0    0  228   10    0  194 1544  305    0    7 |    h = 4
    0    0   56    2    0   30  293   78    0    1 |    i = 4.5
    0    0   64    1    0   29  164   75    0    7 |    j = 5
```

=== Run information ===

Scheme: weka.classifiers.bayes.NaiveBayes

Relation: attributeresult-weka.filters.unsupervised.attribute.NumericToNominal-R4

Instances: 9053

Attributes: 8

- movieId
- imdbId
- tmdbId
- average_rating
- SeasonReleased
- simplifiedPopularity
- profitability_numeric

popularity_score

Test mode: 10-fold cross-validation

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

Naive Bayes Classifier

Class

Attribute 0.5 1 1.5 2 2.5 3 3.5 4 4.5 5

(0.01) (0.02) (0.02) (0.07) (0.09) (0.21) (0.23) (0.25) (0.05) (0.04)

==========

movieId

mean 43114.0652 31728.5233 61683.2062 38383.146 32516.9983 27522.8282 24993.5775 28747.4614 34546.0901 50022.8165

std. dev. 39858.0932 44052.5081 50371.1341 44683.755 41349.8466 38125.996 36432.4383 38352.7922 39779.818 50896.4714

weight sum 93 211 211 640 854 1899 2057 2288 460 340

precision 18.1118 18.1118 18.1118 18.1118 18.1118 18.1118 18.1118 18.1118 18.1118 18.1118

imdbId

mean 551575.3086 574088.8521 1026716.054 632504.6909 514687.1795 440953.3894 390960.9179 438107.6797 478897.3953 570542.9066

std. dev. 809009.0773 892862.4081 1022190.6847 816429.6621 737550.8812 677060.5405 638189.0275 702674.1148 798591.5249 903914.6618

weight sum 93 211 211 640 854 1899 2057 2288 460 340

precision	640.1181	640.1181	640.1181	640.1181	640.1181	640.1181	640.1181	640.1181	640.1181	640.1181

tmdbId

mean	51512.9535	52445.4154	82904.0486	50251.7076	39818.2427	35418.7157	29548.8936	36265.1144	40992.0327	54627.9349
std. dev.	80670.7308	77190.3012	95896.655	69558.8814	62476.9864	56931.1663	52557.1259	58783.3615	61226.9475	71746.0677
weight sum	93	211	211	640	854	1899	2057	2288	460	340
precision	46.0048	46.0048	46.0048	46.0048	46.0048	46.0048	46.0048	46.0048	46.0048	46.0048

SeasonReleased

mean	6.4946	6.6114	6.4408	6.9515	6.6741	6.7155	6.9198	6.748	6.5261	6.7382
std. dev.	3.597	3.5346	3.6804	3.4653	3.5552	3.5059	3.5267	3.5023	3.4831	3.6677
weight sum	93	211	211	639	853	1898	2057	2286	460	340
precision	1	1	1	1	1	1	1	1	1	1

simplifiedPopularity

mean	4.6862	4.9511	3.8728	4.9971	4.9382	5.0692	4.9961	5.0225	4.4655	4.6937
std. dev.	3.8626	4.4579	3.9391	4.4384	4.7013	5.9958	4.8228	4.8872	4.1562	4.2306
weight sum	93	211	211	640	854	1899	2057	2288	460	340
precision	3.2045	3.2045	3.2045	3.2045	3.2045	3.2045	3.2045	3.2045	3.2045	3.2045

profitability_numeric

mean	-0.7419	-0.7156	-0.7156	-0.7375	-0.692	-0.7004	-0.6981	-0.6984	-0.7674	-0.7676
std. dev.	0.4376	0.4511	0.4511	0.44	0.4617	0.4581	0.4591	0.4589	0.4225	0.4223
weight sum	93	211	211	640	854	1899	2057	2288	460	340
precision	1	1	1	1	1	1	1	1	1	1

popularity_score

mean	0.0927	0.1224	0.0644	0.1281	0.1221	0.1339	0.1497	0.153	0.103	0.0944
std. dev.	0.1722	0.3835	0.1487	0.3851	0.4465	0.4321	0.5129	0.4892	0.3421	0.3376
weight sum	93	211	211	640	854	1899	2057	2288	460	340
precision	0.0323	0.0323	0.0323	0.0323	0.0323	0.0323	0.0323	0.0323	0.0323	0.0323

Time taken to build model: 0.01 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	2049	22.6334 %
Incorrectly Classified Instances	7004	77.3666 %
Kappa statistic	0.0266	
Mean absolute error	0.1641	
Root mean squared error	0.2974	
Relative absolute error	99.877 %	
Root relative squared error	103.7542 %	
Total Number of Instances	9053	

=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0.000	0.000	?	0.000	?	?	0.550	0.013	0.5
0.000	0.000	?	0.000	?	?	0.446	0.021	1
0.351	0.107	0.072	0.351	0.120	0.116	0.674	0.054	1.5
0.006	0.002	0.160	0.006	0.012	0.018	0.519	0.081	2
0.000	0.000	?	0.000	?	?	0.505	0.096	2.5
0.085	0.081	0.218	0.085	0.123	0.006	0.510	0.214	3
0.728	0.645	0.249	0.728	0.371	0.073	0.557	0.256	3.5
0.133	0.135	0.250	0.133	0.174	-0.003	0.511	0.261	4
0.000	0.000	?	0.000	?	?	0.531	0.056	4.5
0.021	0.003	0.241	0.021	0.038	0.061	0.592	0.080	5
Weighted Avg. 0.226	0.201	?	0.226	?	?	0.528	0.192	

=== Confusion Matrix ===

a	b	c	d	e	f	g	h	i	j	<-- classified as
0	0	11	0	0	4	54	24	0	0	| a = 0.5
0	0	29	0	0	19	138	24	0	1	| b = 1
0	0	74	1	0	14	85	37	0	0	| c = 1.5
0	0	117	4	0	50	384	83	0	2	| d = 2
0	0	105	2	0	71	554	120	0	2	| e = 2.5
0	0	184	3	0	162	1299	248	0	3	| f = 3
0	0	154	2	0	171	1497	227	0	6	| g = 3.5
0	0	228	10	0	194	1544	305	0	7	| h = 4
0	0	56	2	0	30	293	78	0	1	| i = 4.5
0	0	64	1	0	29	164	75	0	7	| j = 5

3. ADABOOST M1 CLASSIFICATION:

Classifier output

```
=== Run information ===

Scheme:       weka.classifiers.meta.AdaBoostM1 -P 100 -S 1 -I 10 -W weka.classifiers.trees.DecisionStum
Relation:     attributeresult-weka.filters.unsupervised.attribute.NumericToNominal-R4
Instances:    9053
Attributes:   8
              movieId
              imdbId
              tmdbId
              average_rating
              SeasonReleased
              simplifiedPopularity
              profitability_numeric
              popularity_score
Test mode:    10-fold cross-validation

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

AdaBoostM1: No boosting possible, one classifier used!
Decision Stump

Classifications

tmdbId <= 1246.5 : 4
tmdbId > 1246.5 : 4
tmdbId is missing : 4

Class distributions

tmdbId <= 1246.5
0.5     1       1.5     2       2.5     3       3.5     4       4.5     5
0.004103967168262654    0.005471956224350205    0.0027359781121751026   0.015047879616963064    0.01915
```

Classifier output

```
Class distributions

tmdbId <= 1246.5
0.5     1       1.5     2       2.5     3       3.5     4       4.5     5
0.004103967168262654    0.005471956224350205    0.0027359781121751026   0.015047879616963064    0.01915
tmdbId > 1246.5
0.5     1       1.5     2       2.5     3       3.5     4       4.5     5
0.010814708002883922    0.02487382840663302     0.02511415525114155     0.07558279259793318     0.10093
tmdbId is missing
0.5     1       1.5     2       2.5     3       3.5     4       4.5     5
0.010272837733348061    0.023307190986413345    0.023307190986413345    0.07069479730476086     0.09433



Time taken to build model: 0.03 seconds

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

Correctly Classified Instances        2292               25.3176 %
Incorrectly Classified Instances      6761               74.6824 %
Kappa statistic                          0.0019
Mean absolute error                      0.1635
Root mean squared error                  0.286
Relative absolute error                 99.5181 %
Root relative squared error             99.8045 %
Total Number of Instances             9053

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.000    0.000    ?          0.000    ?          ?        0.569     0.012     0.5
                 0.000    0.000    ?          0.000    ?          ?        0.533     0.025     1
```

```
=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.000    0.000    ?          0.000    ?          ?        0.569     0.012     0.5
                 0.000    0.000    ?          0.000    ?          ?        0.533     0.025     1
                 0.000    0.000    ?          0.000    ?          ?        0.563     0.027     1.5
                 0.000    0.000    ?          0.000    ?          ?        0.529     0.074     2
                 0.000    0.000    ?          0.000    ?          ?        0.525     0.099     2.5
                 0.000    0.000    ?          0.000    ?          ?        0.516     0.215     3
                 0.047    0.037    0.272      0.047    0.080      0.022    0.521     0.242     3.5
                 0.959    0.961    0.252      0.959    0.400      -0.004   0.540     0.291     4
                 0.000    0.000    ?          0.000    ?          ?        0.526     0.056     4.5
                 0.000    0.000    ?          0.000    ?          ?        0.546     0.042     5
Weighted Avg.    0.253    0.251    ?          0.253    ?          ?        0.529     0.194

=== Confusion Matrix ===

    a    b    c    d    e    f    g    h    i    j   <-- classified as
    0    0    0    0    0    0    3   90    0    0 |    a = 0.5
    0    0    0    0    0    0    6  205    0    0 |    b = 1
    0    0    0    0    0    0    4  207    0    0 |    c = 1.5
    0    0    0    0    0    0   27  613    0    0 |    d = 2
    0    0    0    0    0    0   32  822    0    0 |    e = 2.5
    0    0    0    0    0    0   74 1825    0    0 |    f = 3
    0    0    0    0    0    0   97 1960    0    0 |    g = 3.5
    0    0    0    0    0    0   93 2195    0    0 |    h = 4
    0    0    0    0    0    0   14  446    0    0 |    i = 4.5
    0    0    0    0    0    0    7  333    0    0 |    j = 5
```

=== Run information ===

Scheme: weka.classifiers.meta.AdaBoostM1 -P 100 -S 1 -I 10 -W weka.classifiers.trees.DecisionStump

Relation: attributeresult-weka.filters.unsupervised.attribute.NumericToNominal-R4

Instances: 9053

Attributes: 8

movieId

imdbId

tmdbId

average_rating

SeasonReleased

simplifiedPopularity

profitability_numeric

popularity_score

Test mode: 10-fold cross-validation

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

AdaBoostM1: No boosting possible, one classifier used!

Decision Stump

Classifications

tmdbId <= 1246.5 : 4

tmdbId > 1246.5 : 4

tmdbId is missing : 4

Class distributions

tmdbId <= 1246.5

0.5 1 1.5 2 2.5 3 3.5 4 4.5 5

0.004103967168262654 0.00547195622435020 5 0.0027359781121751026 0.015047879616963064 0.019151846785225718 0.1409028727770178

0.3023255813953488 0.43091655266757867 0.06839945280437756
0.01094391244870041

tmdbId > 1246.5

0.5 1 1.5 2 2.5 3 3.5 4 4.5 5

0.010814708002883922 0.02487382840663302 0.02511415525114155
0.07558279259793318 0.10093727469358327 0.21581350636866137
0.220620043258832 0.2370824321076664 0.049267003124248976
0.039894256188416244

tmdbId is missing

0.5 1 1.5 2 2.5 3 3.5 4 4.5 5

0.010272837733348061 0.023307190986413345 0.023307190986413345
0.07069479730476086 0.09433337015354026 0.20976471887772008
0.22721749696233293 0.25273390036452004 0.050811885562796866
0.0375566110681542

Time taken to build model: 0.03 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	2292	25.3176 %
Incorrectly Classified Instances	6761	74.6824 %
Kappa statistic	0.0019	
Mean absolute error	0.1635	
Root mean squared error	0.286	
Relative absolute error	99.5181 %	
Root relative squared error	99.8045 %	
Total Number of Instances	9053	

=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0.000	0.000	?	0.000	?	?	0.569	0.012	0.5
0.000	0.000	?	0.000	?	?	0.533	0.025	1
0.000	0.000	?	0.000	?	?	0.563	0.027	1.5
0.000	0.000	?	0.000	?	?	0.529	0.074	2
0.000	0.000	?	0.000	?	?	0.525	0.099	2.5
0.000	0.000	?	0.000	?	?	0.516	0.215	3
0.047	0.037	0.272	0.047	0.080	0.022	0.521	0.242	3.5
0.959	0.961	0.252	0.959	0.400	-0.004	0.540	0.291	4
0.000	0.000	?	0.000	?	?	0.526	0.056	4.5
0.000	0.000	?	0.000	?	?	0.546	0.042	5
Weighted Avg. 0.253	0.251	?	0.253	?	?	0.529	0.194	

=== Confusion Matrix ===

a	b	c	d	e	f	g	h	i	j	<-- classified as
0	0	0	0	0	0	3	90	0	0	| a = 0.5
0	0	0	0	0	0	6	205	0	0	| b = 1
0	0	0	0	0	0	4	207	0	0	| c = 1.5
0	0	0	0	0	0	27	613	0	0	| d = 2
0	0	0	0	0	0	32	822	0	0	| e = 2.5
0	0	0	0	0	0	74	1825	0	0	| f = 3
0	0	0	0	0	0	97	1960	0	0	| g = 3.5
0	0	0	0	0	0	93	2195	0	0	| h = 4
0	0	0	0	0	0	14	446	0	0	| i = 4.5
0	0	0	0	0	0	7	333	0	0	| j = 5

4. RANDOM FOREST CLASSIFICATION:

Classifier output

```
=== Run information ===

Scheme:       weka.classifiers.trees.RandomForest -P 100 -I 100 -num-slots 1 -K 0 -M 1.0 -V 0.001 -S 1
Relation:     attributeresult-weka.filters.unsupervised.attribute.NumericToNominal-R4
Instances:    9053
Attributes:   8
              movieId
              imdbId
              tmdbId
              average_rating
              SeasonReleased
              simplifiedPopularity
              profitability_numeric
              popularity_score
Test mode:    10-fold cross-validation

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

RandomForest

Bagging with 100 iterations and base learner

weka.classifiers.trees.RandomTree -K 0 -M 1.0 -V 0.001 -S 1 -do-not-check-capabilities

Time taken to build model: 3.01 seconds
```

Classifier output

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

Correctly Classified Instances        2322               25.649  %
Incorrectly Classified Instances      6731               74.351  %
Kappa statistic                          0.0546
Mean absolute error                      0.1597
Root mean squared error                  0.2892
Relative absolute error                 97.1729 %
Root relative squared error            100.8961 %
Total Number of Instances             9053

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.000    0.001    0.000      0.000    0.000      -0.003   0.674     0.018     0.5
                 0.005    0.005    0.021      0.005    0.008      -0.001   0.556     0.030     1
                 0.024    0.006    0.089      0.024    0.037      0.035    0.602     0.036     1.5
                 0.048    0.031    0.107      0.048    0.067      0.026    0.596     0.095     2
                 0.064    0.045    0.129      0.064    0.086      0.026    0.580     0.123     2.5
                 0.298    0.241    0.247      0.298    0.270      0.054    0.563     0.254     3
                 0.324    0.261    0.267      0.324    0.293      0.059    0.575     0.275     3.5
                 0.420    0.334    0.298      0.420    0.349      0.078    0.580     0.336     4
                 0.009    0.010    0.043      0.009    0.014      -0.004   0.607     0.072     4.5
                 0.097    0.011    0.260      0.097    0.141      0.139    0.682     0.124     5
Weighted Avg.    0.256    0.202    0.222      0.256    0.231      0.054    0.583     0.229
```

```
=== Confusion Matrix ===

   a   b   c   d   e   f   g   h   i   j   <-- classified as
   0   0   0   4   5  23  16  44   0   1 |   a = 0.5
   0   1   1  10  11  70  52  64   0   2 |   b = 1
   0   2   5  20  13  55  39  70   4   3 |   c = 1.5
   2   4   8  31  46 174 172 196   2   5 |   d = 2
   1   6   9  33  55 260 220 252   8  10 |   e = 2.5
   2  11  14  65  96 566 539 567  18  21 |   f = 3
   2  12   4  55  89 500 666 703  19   7 |   g = 3.5
   3   6  10  53  86 490 618 961  30  31 |   h = 4
   0   2   3   7  14  83 123 210   4  14 |   i = 4.5
   0   3   2  11  13  68  49 153   8  33 |   j = 5
```

=== Run information ===

Scheme: weka.classifiers.trees.RandomForest -P 100 -I 100 -num-slots 1 -K 0 -M 1.0 -V 0.001 -S 1

Relation: attributeresult-weka.filters.unsupervised.attribute.NumericToNominal-R4

Instances: 9053

Attributes: 8

movieId

imdbId

tmdbId

average_rating

SeasonReleased

simplifiedPopularity

profitability_numeric

popularity_score

Test mode: 10-fold cross-validation

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

RandomForest

Bagging with 100 iterations and base learner

weka.classifiers.trees.RandomTree -K 0 -M 1.0 -V 0.001 -S 1 -do-not-check-capabilities

Time taken to build model: 3.01 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	2322	25.649 %
Incorrectly Classified Instances	6731	74.351 %
Kappa statistic	0.0546	
Mean absolute error	0.1597	
Root mean squared error	0.2892	
Relative absolute error	97.1729 %	
Root relative squared error	100.8961 %	
Total Number of Instances	9053	

=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0.000	0.001	0.000	0.000	0.000	-0.003	0.674	0.018	0.5
0.005	0.005	0.021	0.005	0.008	-0.001	0.556	0.030	1
0.024	0.006	0.089	0.024	0.037	0.035	0.602	0.036	1.5
0.048	0.031	0.107	0.048	0.067	0.026	0.596	0.095	2
0.064	0.045	0.129	0.064	0.086	0.026	0.580	0.123	2.5
0.298	0.241	0.247	0.298	0.270	0.054	0.563	0.254	3
0.324	0.261	0.267	0.324	0.293	0.059	0.575	0.275	3.5

	0.420	0.334	0.298	0.420	0.349	0.078	0.580	0.336	4
	0.009	0.010	0.043	0.009	0.014	-0.004	0.607	0.072	4.5
	0.097	0.011	0.260	0.097	0.141	0.139	0.682	0.124	5
Weighted Avg.	0.256	0.202	0.222	0.256	0.231	0.054	0.583	0.229	

=== Confusion Matrix ===

a	b	c	d	e	f	g	h	i	j	<-- classified as
0	0	0	4	5	23	16	44	0	1	a = 0.5
0	1	1	10	11	70	52	64	0	2	b = 1
0	2	5	20	13	55	39	70	4	3	c = 1.5
2	4	8	31	46	174	172	196	2	5	d = 2
1	6	9	33	55	260	220	252	8	10	e = 2.5
2	11	14	65	96	566	539	567	18	21	f = 3
2	12	4	55	89	500	666	703	19	7	g = 3.5
3	6	10	53	86	490	618	961	30	31	h = 4
0	2	3	7	14	83	123	210	4	14	i = 4.5
0	3	2	11	13	68	49	153	8	33	j = 5

5. DECISION TREE CLASSIFICATION:

Classifier output

```
=== Run information ===

Scheme:       weka.classifiers.rules.DecisionTable -X 1 -S "weka.attributeSelection.BestFirst -D 1 -N 5
Relation:     attributeresult-weka.filters.unsupervised.attribute.NumericToNominal-R4
Instances:    9053
Attributes:   8
              movieId
              imdbId
              tmdbId
              average_rating
              SeasonReleased
              simplifiedPopularity
              profitability_numeric
              popularity_score
Test mode:    10-fold cross-validation

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

Decision Table:

Number of training instances: 9053
Number of Rules : 15
Non matches covered by Majority class.
        Best first.
        Start set: no attributes
        Search direction: forward
        Stale search after 5 node expansions
        Total number of subsets evaluated: 32
        Merit of best subset found:   27.483
Evaluation (for feature selection): CV (leave one out)
Feature set: 2,3,5,4
```

Classifier output

```
Time taken to build model: 0.26 seconds

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

Correctly Classified Instances        2469               27.2727 %
Incorrectly Classified Instances      6584               72.7273 %
Kappa statistic                          0.0491
Mean absolute error                      0.1621
Root mean squared error                  0.2847
Relative absolute error                 98.6353 %
Root relative squared error             99.3344 %
Total Number of Instances             9053

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.000    0.000    ?          0.000    ?          ?        0.685     0.017     0.5
                 0.000    0.000    ?          0.000    ?          ?        0.607     0.031     1
                 0.000    0.000    ?          0.000    ?          ?        0.682     0.048     1.5
                 0.000    0.000    0.000      0.000    0.000      -0.004   0.596     0.091     2
                 0.000    0.002    0.000      0.000    0.000      -0.012   0.566     0.106     2.5
                 0.291    0.243    0.241      0.291    0.264      0.045    0.558     0.238     3
                 0.243    0.171    0.294      0.243    0.266      0.077    0.580     0.279     3.5
                 0.611    0.530    0.281      0.611    0.384      0.071    0.579     0.327     4
                 0.000    0.000    ?          0.000    ?          ?        0.636     0.084     4.5
                 0.059    0.006    0.290      0.059    0.098      0.116    0.676     0.091     5
Weighted Avg.    0.273    0.224    ?          0.273    ?          ?        0.586     0.222
```

```
=== Confusion Matrix ===

    a    b    c    d    e    f    g    h    i    j   <-- classified as
    0    0    0    0    0   18    4   71    0    0 |    a = 0.5
    0    0    0    0    0   60   26  124    0    1 |    b = 1
    0    0    0    0    1   41   16  151    0    2 |    c = 1.5
    0    0    0    0    1  182  111  343    0    3 |    d = 2
    0    0    0    0    0  255  193  401    0    5 |    e = 2.5
    0    0    0    0    3  553  425  908    0   10 |    f = 3
    0    0    0    2    4  527  499 1013    0   12 |    g = 3.5
    0    0    0    0    2  522  356 1397    0   11 |    h = 4
    0    0    0    0    0   75   45  335    0    5 |    i = 4.5
    0    0    0    0    2   60   21  237    0   20 |    j = 5
```

=== Run information ===

Scheme: weka.classifiers.rules.DecisionTable -X 1 -S "weka.attributeSelection.BestFirst -D 1 -N 5"

Relation: attributeresult-weka.filters.unsupervised.attribute.NumericToNominal-R4

Instances: 9053

Attributes: 8

movieId

imdbId

tmdbId

average_rating

SeasonReleased

simplifiedPopularity

profitability_numeric

popularity_score

Test mode: 10-fold cross-validation

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

Decision Table:

Number of training instances: 9053

Number of Rules : 15

Non matches covered by Majority class.

Best first.

Start set: no attributes

Search direction: forward

Stale search after 5 node expansions

Total number of subsets evaluated: 32

Merit of best subset found: 27.483

Evaluation (for feature selection): CV (leave one out)

Feature set: 2,3,5,4

Time taken to build model: 0.26 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	2469	27.2727 %
Incorrectly Classified Instances	6584	72.7273 %
Kappa statistic	0.0491	
Mean absolute error	0.1621	
Root mean squared error	0.2847	
Relative absolute error	98.6353 %	
Root relative squared error	99.3344 %	
Total Number of Instances	9053	

=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0.000	0.000	?	0.000	?	?	0.685	0.017	0.5
0.000	0.000	?	0.000	?	?	0.607	0.031	1
0.000	0.000	?	0.000	?	?	0.682	0.048	1.5
0.000	0.000	0.000	0.000	0.000	-0.004	0.596	0.091	2
0.000	0.002	0.000	0.000	0.000	-0.012	0.566	0.106	2.5
0.291	0.243	0.241	0.291	0.264	0.045	0.558	0.238	3
0.243	0.171	0.294	0.243	0.266	0.077	0.580	0.279	3.5
0.611	0.530	0.281	0.611	0.384	0.071	0.579	0.327	4
0.000	0.000	?	0.000	?	?	0.636	0.084	4.5
0.059	0.006	0.290	0.059	0.098	0.116	0.676	0.091	5
Weighted Avg. 0.273	0.224	?	0.273	?	?	0.586	0.222	

=== Confusion Matrix ===

a	b	c	d	e	f	g	h	i	j	<-- classified as
0	0	0	0	0	18	4	71	0	0	a = 0.5
0	0	0	0	0	60	26	124	0	1	b = 1
0	0	0	0	1	41	16	151	0	2	c = 1.5
0	0	0	0	1	182	111	343	0	3	d = 2
0	0	0	0	0	255	193	401	0	5	e = 2.5
0	0	0	0	3	553	425	908	0	10	f = 3
0	0	0	2	4	527	499	1013	0	12	g = 3.5
0	0	0	0	2	522	356	1397	0	11	h = 4
0	0	0	0	0	75	45	335	0	5	i = 4.5
0	0	0	0	2	60	21	237	0	20	j = 5

F-score Results of Classifiers for Original and Extended Dataset

Classification Model	Without Attributes	With Attributes
Bayesnet	0.2	0.202
Naivebayes	0.152	0.832
AdaBoost1	0.048	0.104
RandomForest	0.23	0.231
Decision Tree	0.1012	0.253

It is evident that the classifier constructed from the decision table, incorporating four additional attributes, outperforms the classifier without these attributes, as indicated by a higher F-score.

***The Weighted Average of the F-score is computed by treating '?' as '0'. ***

ANALYSIS OF ACTION RULE MINING USING LISP MINER

Analysis of Action Rule mining using LISP Miner: The objective attribute, average rating, is segmented into 10 distinct categories: 0.5, 1.0, 1.5, 2.0, 2.5, 3.0, 3.5, 4.0, 4.5, and 5.0.

Antecedent Stable: Profitability_numeric

Antecedent Variable: Popularity_score, SeasonReleased, simplifiedPopularity

Succedent Variable: average_rating

The main objective in the field of Action Rule Mining with LISP Miner is to predict the average rating, which is divided into 10 different categories ranging from 0.5 to 5.0. The target variable, "average_rating," is the result of choosing antecedents that include a stable component, "Profitability_numeric," and dynamic features, such as "Popularity_score," "SeasonReleased," and "simplifiedPopularity."

- Stable Antecedent - "Profitability_numeric": It is determined that the characteristic "Profitability_numeric" is the only one that consistently predicts average ratings. This binary variable is a consistent predictor of a film's profitability because it doesn't change over time.
- Variable Antecedents - "Popularity_score," "SeasonReleased," and "simplifiedPopularity": It is understood that certain attributes are changeable or variable, such as "Popularity_score," "SeasonReleased," and "simplifiedPopularity". These characteristics display variations impacted by various outside variables.
- "Popularity_score" is dynamic and responsive to audience responses; it varies based on average ratings and viewer votes.
- "SeasonReleased" is classified as a variable property since it varies with the seasons, which may have an effect on audience preferences and movie performance.
- "simplifiedPopularity" changes according on the general popularity of a film, which is affected by things like advertising tactics and audience interest.

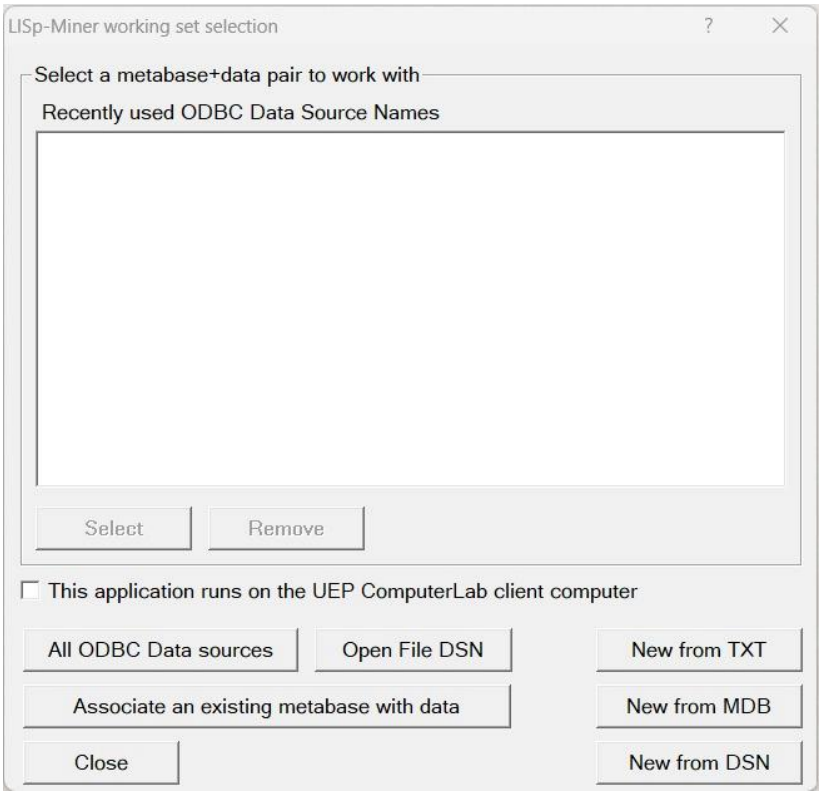
In this action rule mining procedure, the purposeful selection of both stable and variable antecedents shows a strategic decision that balances the dependability and stability of attributes in predicting the average movie rating.

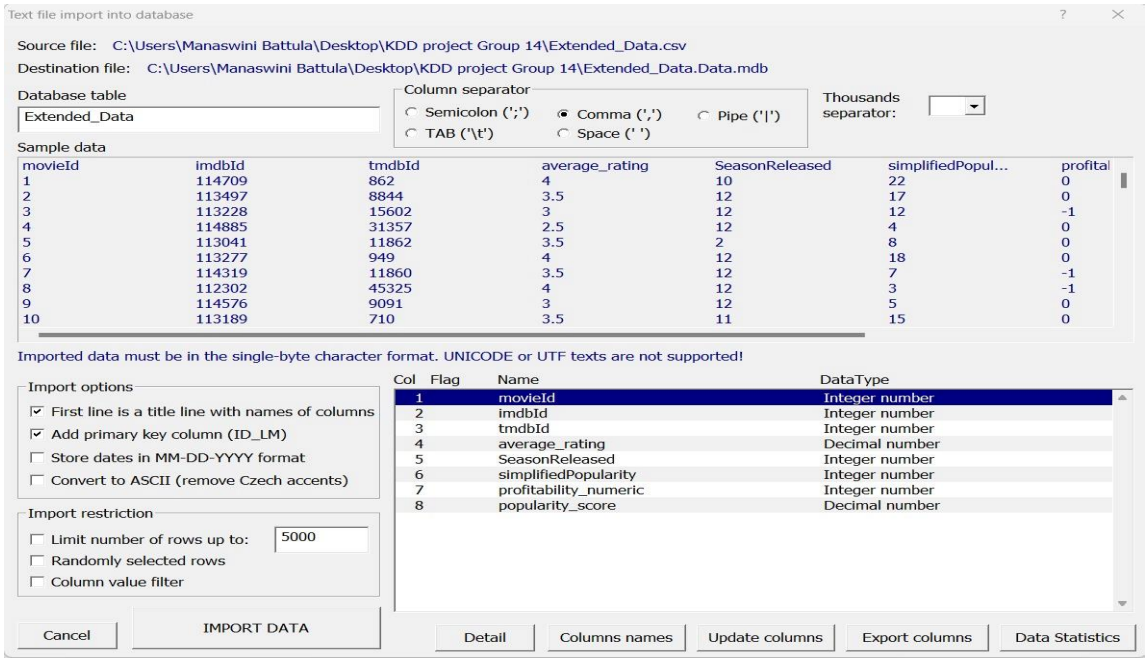
Coefficient_Type: One Category

Action rules outline the potential transformations of objects from one state to another concerning a distinctive attribute. In this context, we will derive Action rules using Lisp Miner software, having already discretized the decision attribute.

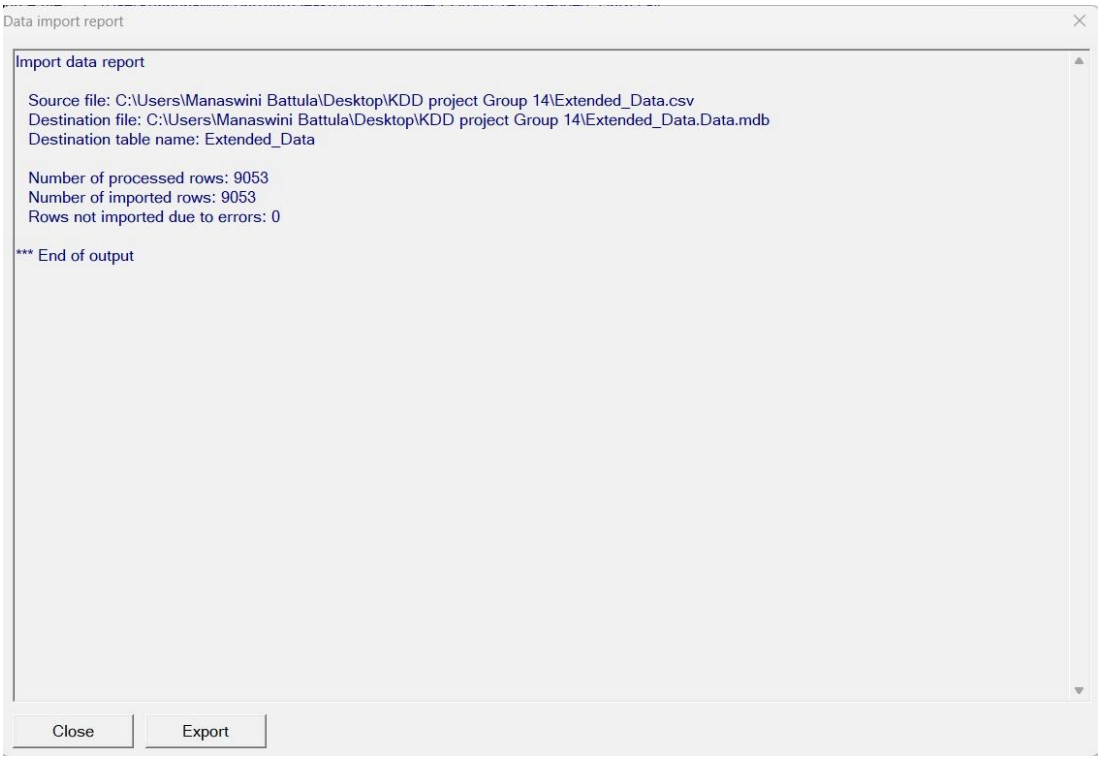
Steps with Screenshots:

1. Load the Extended CSV file into Lisp Miner by selecting "New from TXT" and choosing the Extended.csv file.

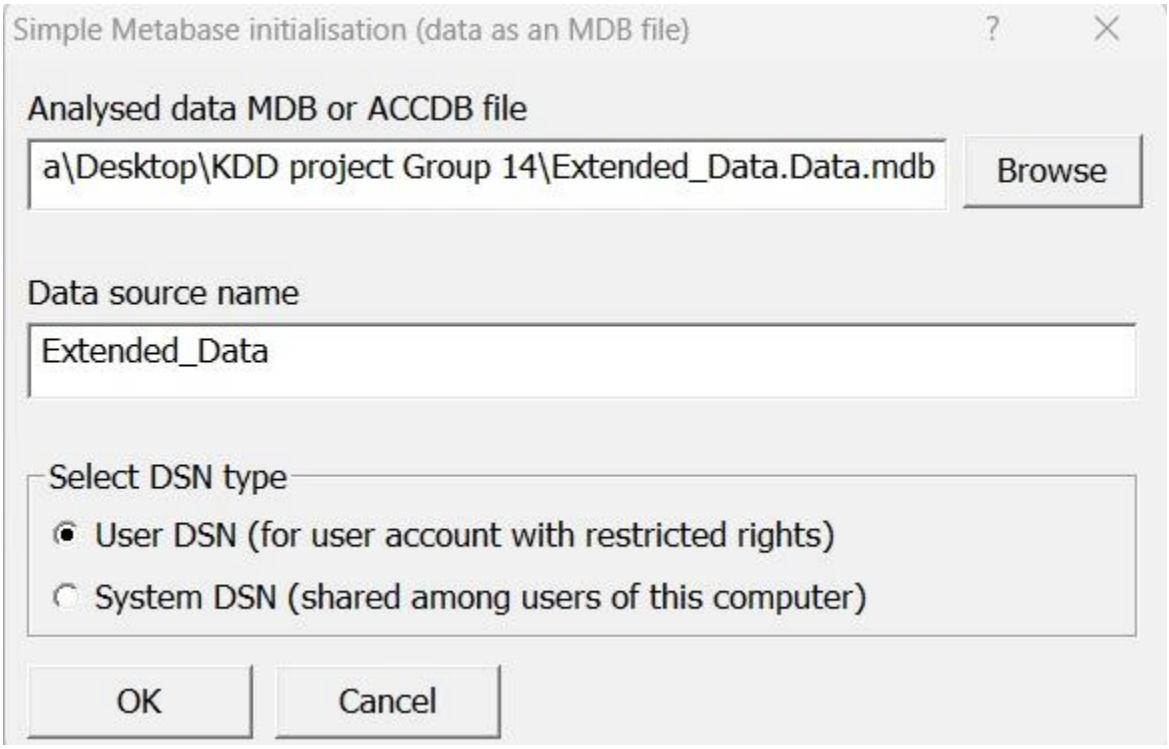




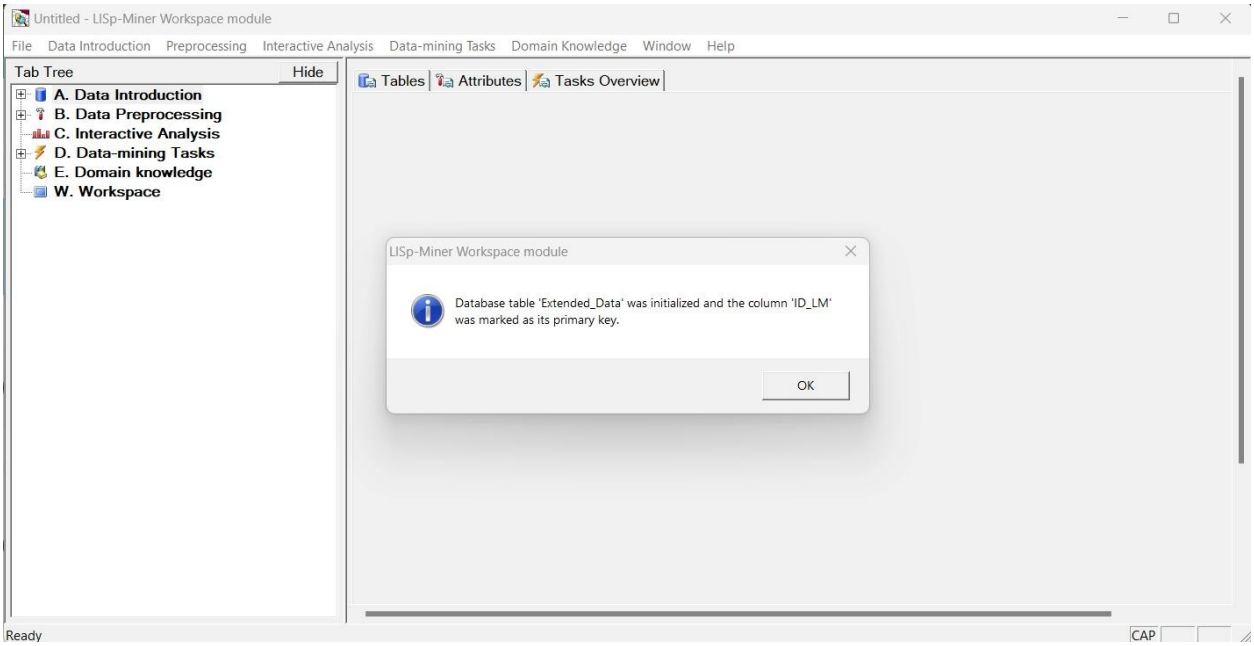
Now, Data report has been imported to the file:

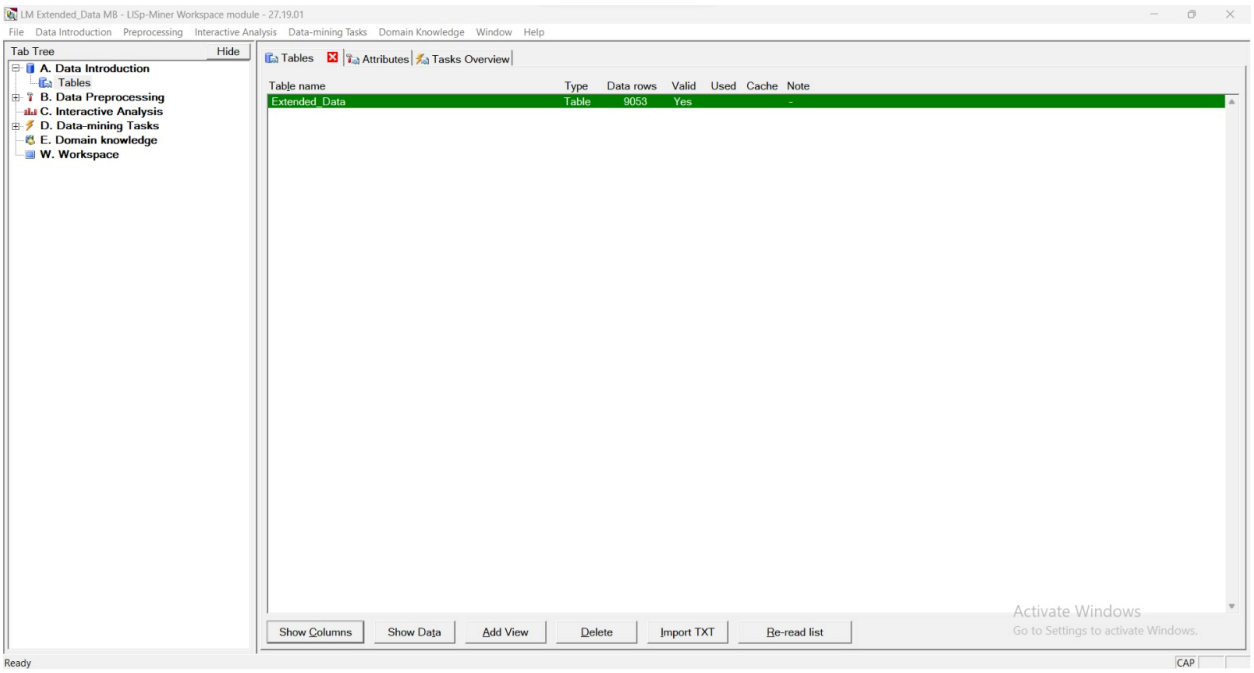


Selected DSN type as User DSN and clicked ok:

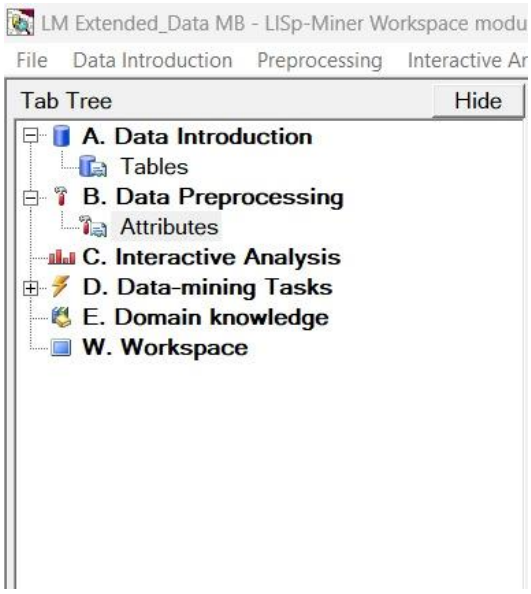


Here, The Database table “Extended Data” is initialized and Column ‘ID_LM’ is marked as its primary key:

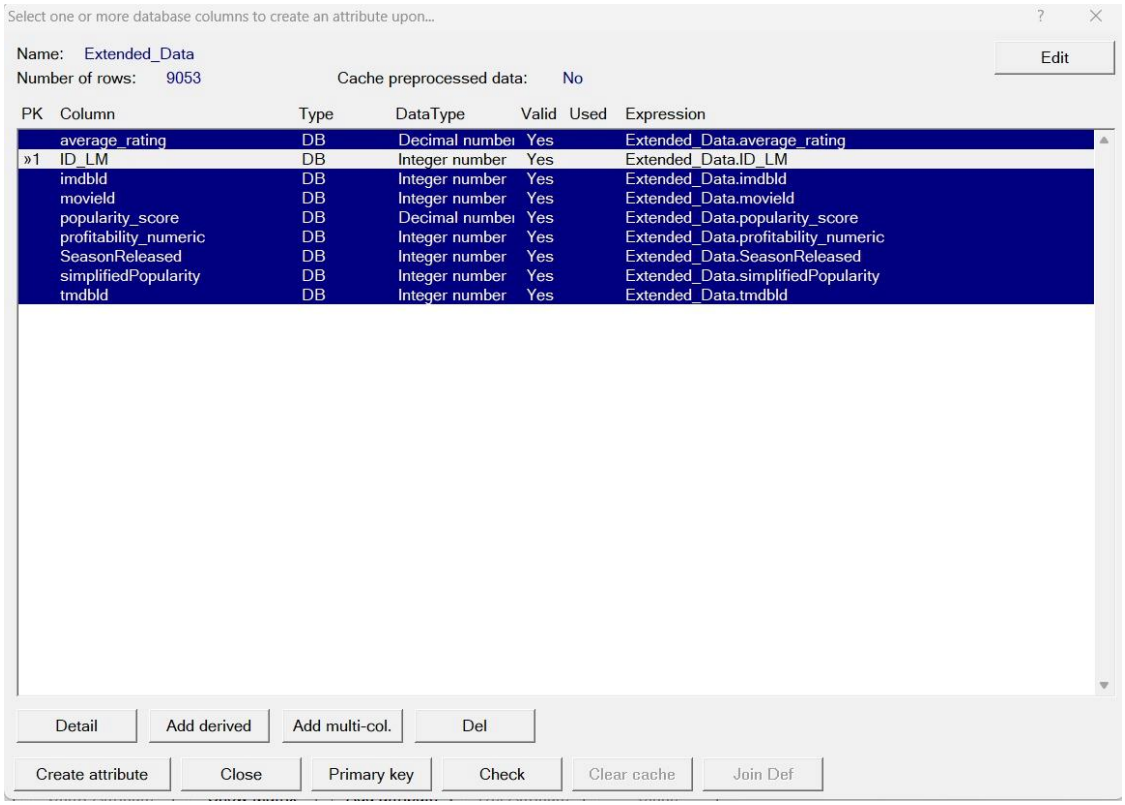




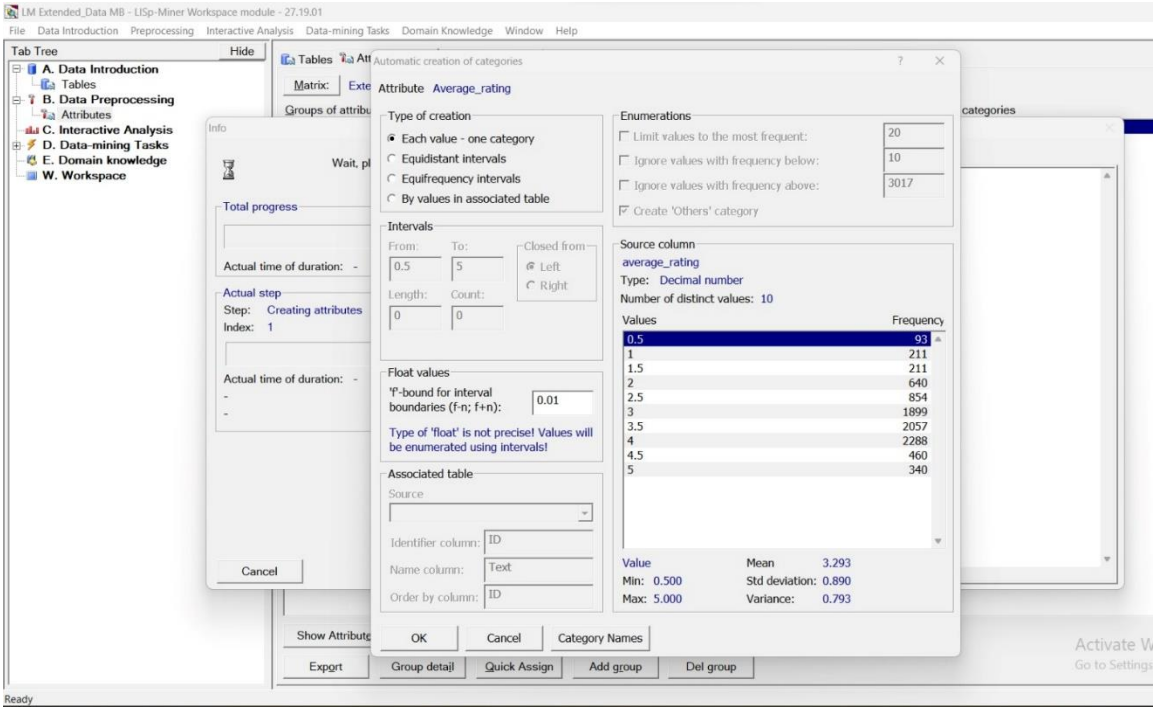
From Tab Tree select the Attributes from the Data Preprocessing and Open:

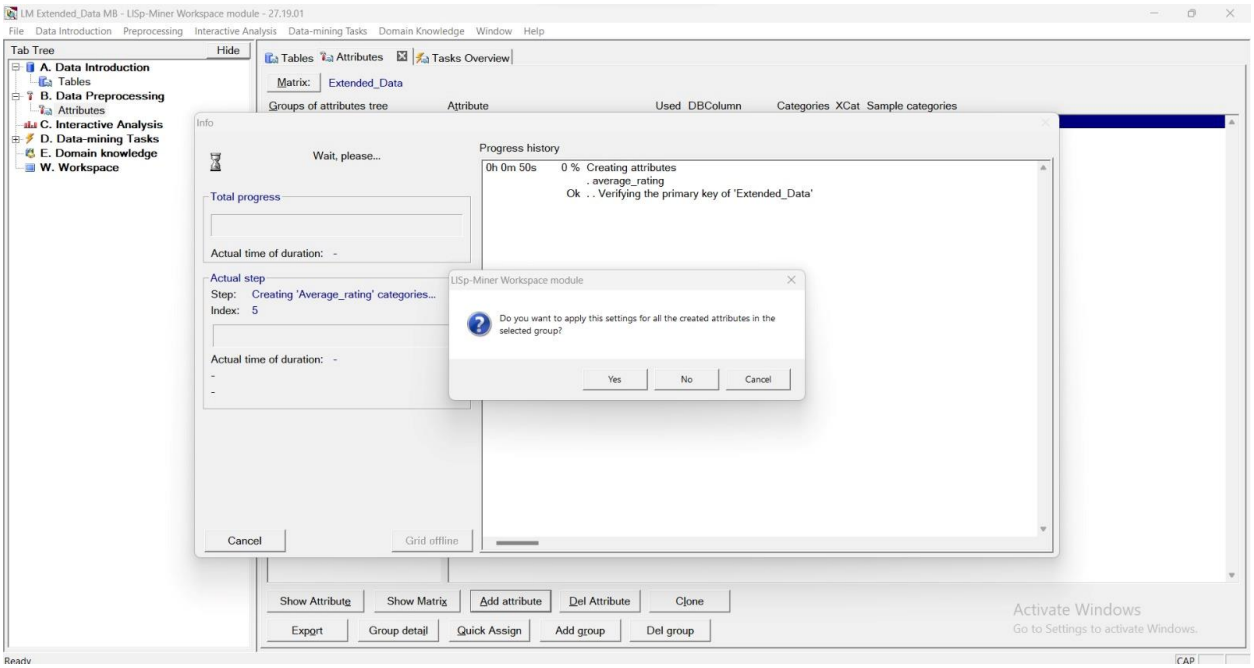


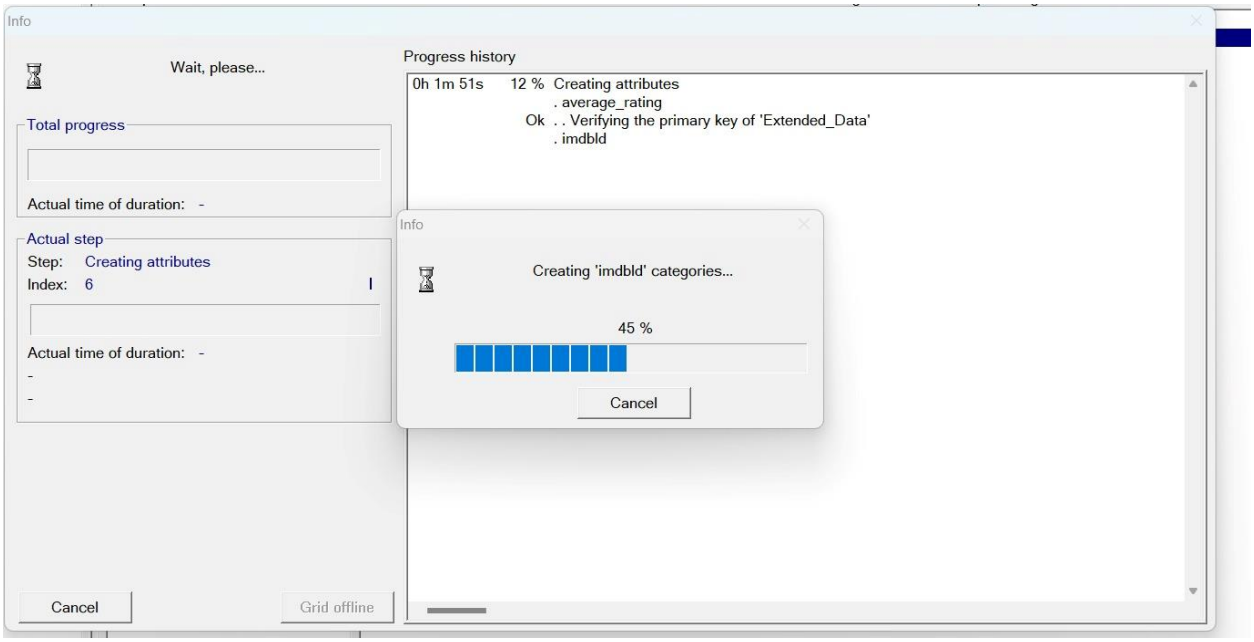
2. Add all attributes except ID_LM



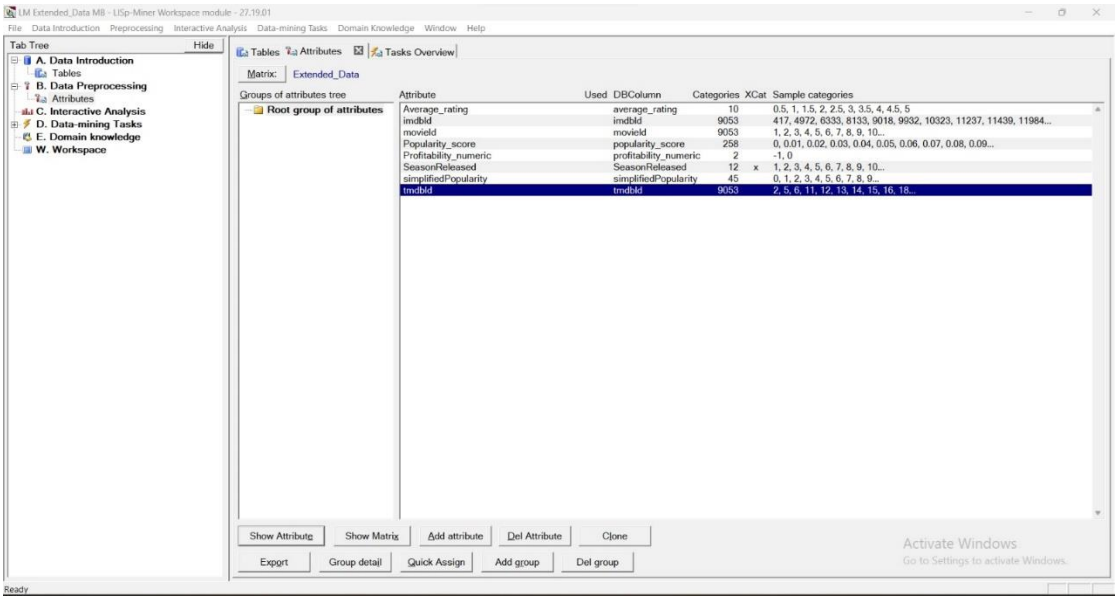
- Applying the settings for the created attributes (all of them):



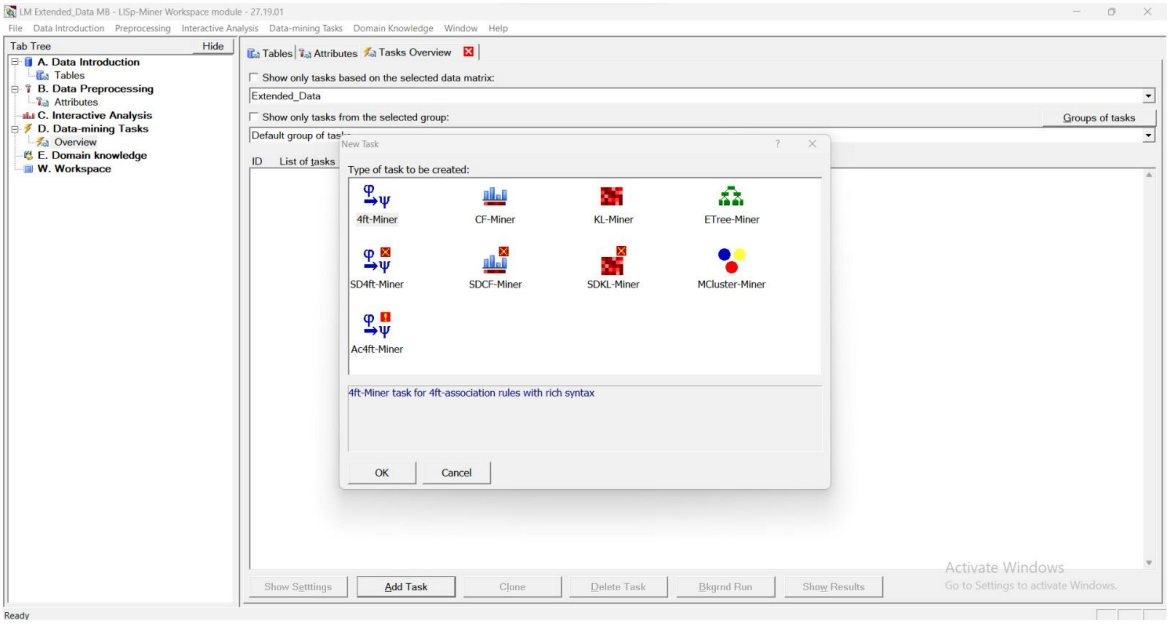

LM Extended_Data MB - LISp-Miner Workspace module - 27.19.01
File Data Introduction Preprocessing Interactive Analysis Data-mining Tasks Domain Knowledge Window Help
Tab Tree Hide
A. Data Introduction
Tables
B. Data Preprocessing
Attributes
C. Interactive Analysis
D. Data-mining Tasks
E. Domain knowledge
W. Workspace
Tables Attributes Tasks Overview
Matrix: Extended_Data
Groups of attributes tree Attribute Used DBColumn Categories XCat Sample categories
Info
Wait, please...
Progress history
0h 0m 50s 0 % Creating attributes
. average_rating
Ok . . Verifying the primary key of 'Extended_Data'
Total progress
Actual time of duration: -
Actual step
Step: Creating 'Average_rating' categories...
Index: 5
Actual time of duration: -
LISp-Miner Workspace module
Do you want to apply this settings for all the created attributes in the selected group?
Yes No Cancel
Cancel Grid offline
Show Attribute Show Matrix Add attribute Del Attribute Clone
Export Group detail Quick Assign Add group Del group
Activate Windows
Go to Settings to activate Windows.
Ready
CAP


Info
Wait, please...
Progress history
0h 1m 51s 12 % Creating attributes
. average_rating
Ok . . Verifying the primary key of 'Extended_Data'
. imdbld
Total progress
Actual time of duration: -
Actual step
Step: Creating attributes
Index: 6
Actual time of duration: -
Info
Creating 'imdbld' categories...
45 %
Cancel
Cancel Grid offline

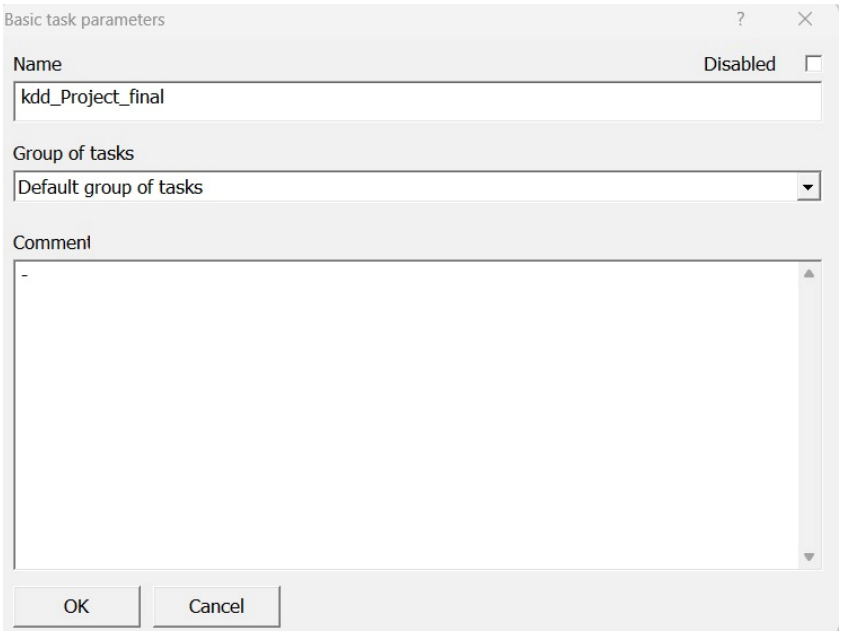
- All the attributes are obtained as follows:



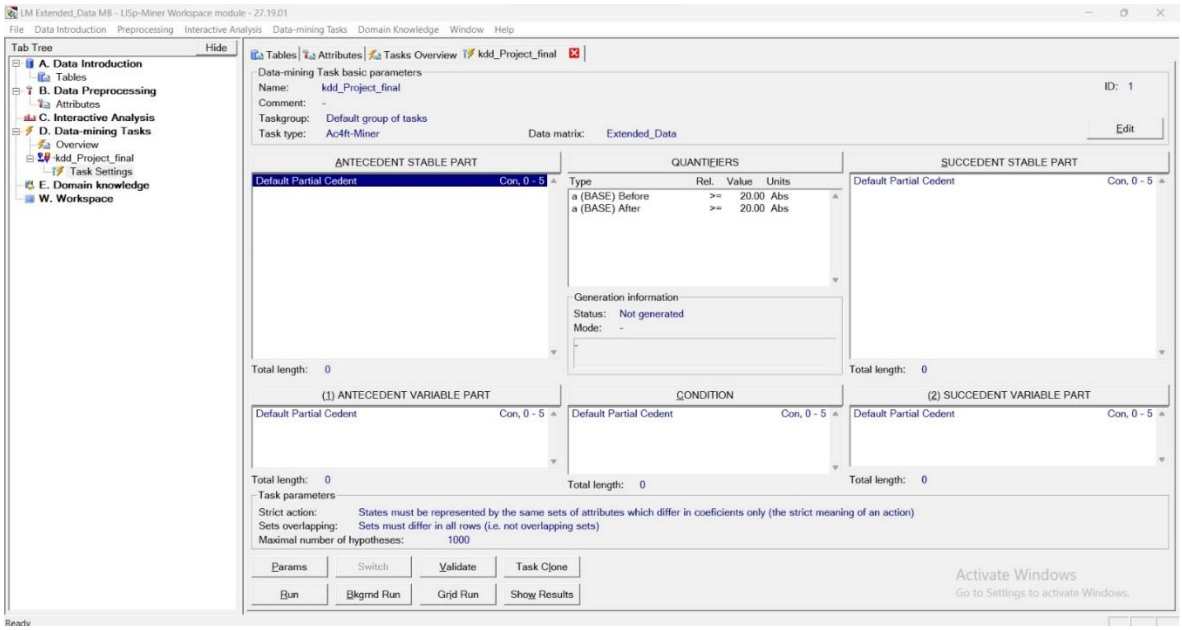
- Select Task Overview and initiate a new AC4ft-Miner Task, assigning it a name.

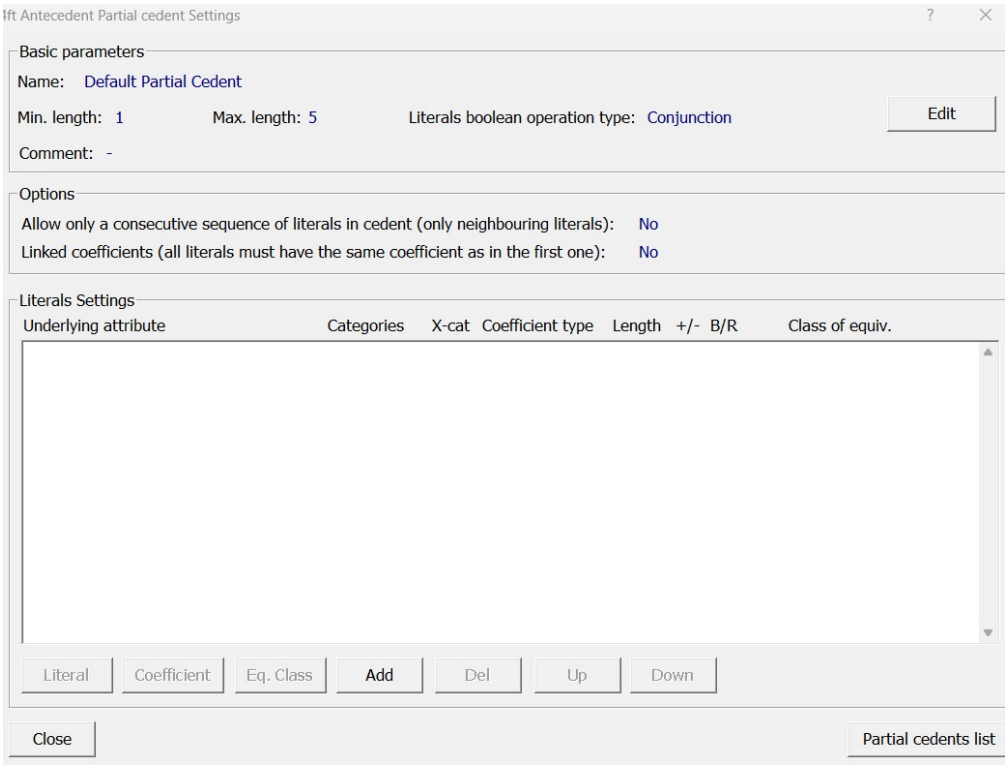


- Now Assigning a group name for the task:

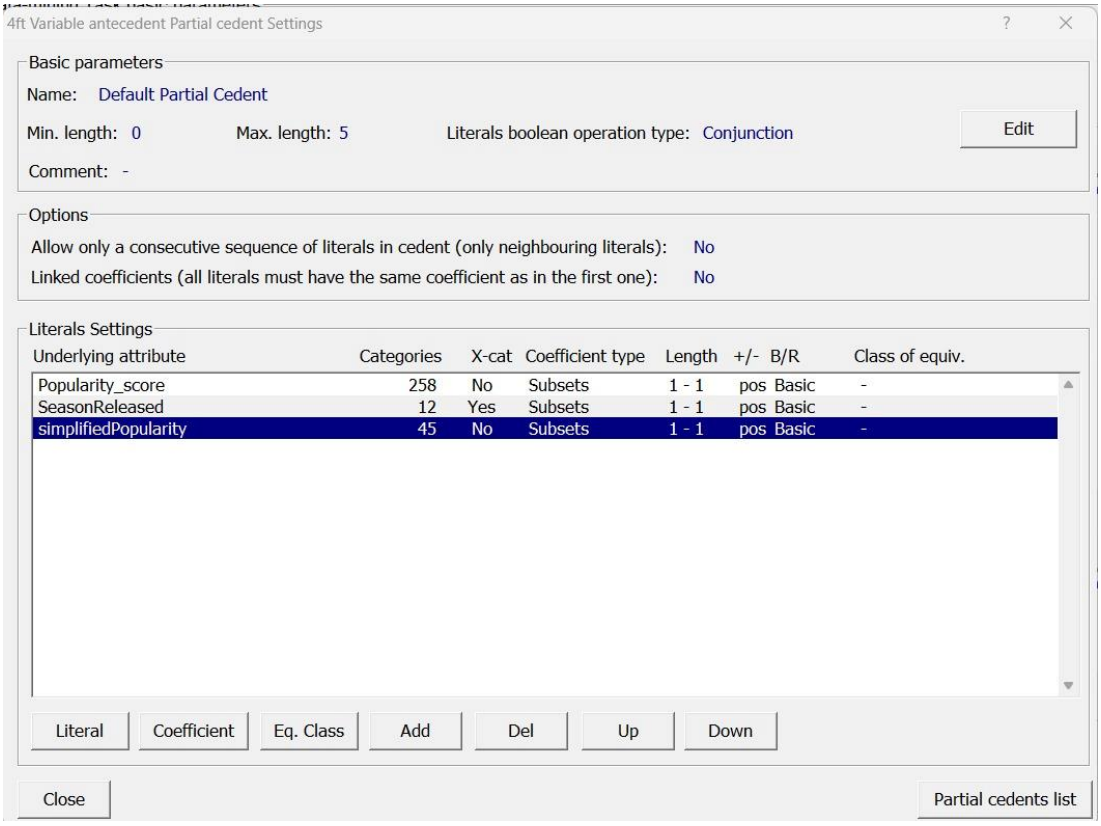


3. Modify the minimum length for both Antecedent and Succedent, adjusting it from 0 to 1. Additionally, allocate attributes to Antecedents and Succedent.

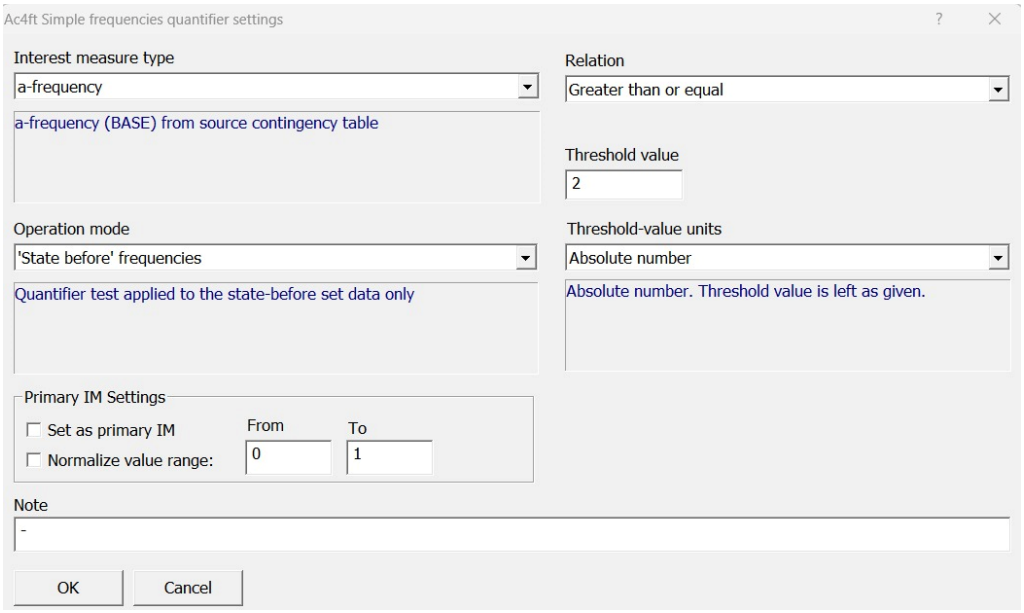




Set first value:

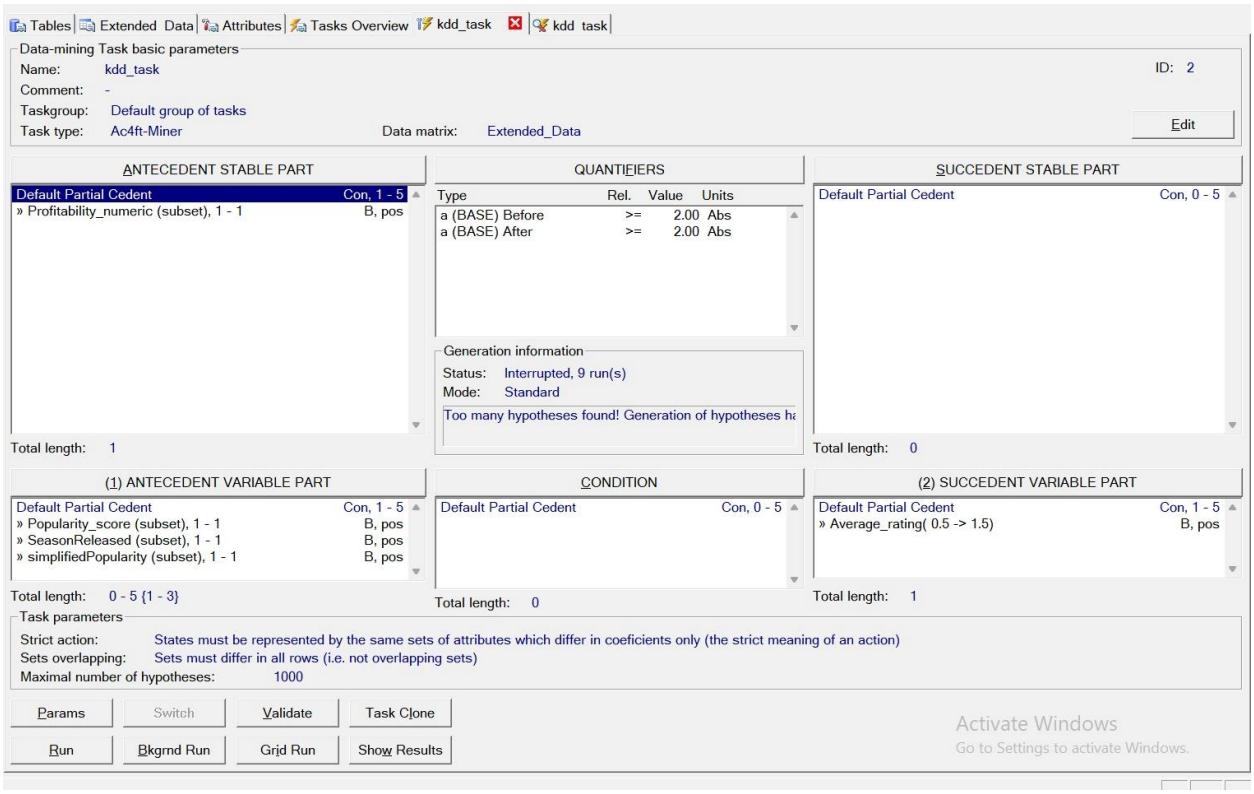


In the below figure- Reducing the threshold value to 2 from 20 in the quantifier settings:

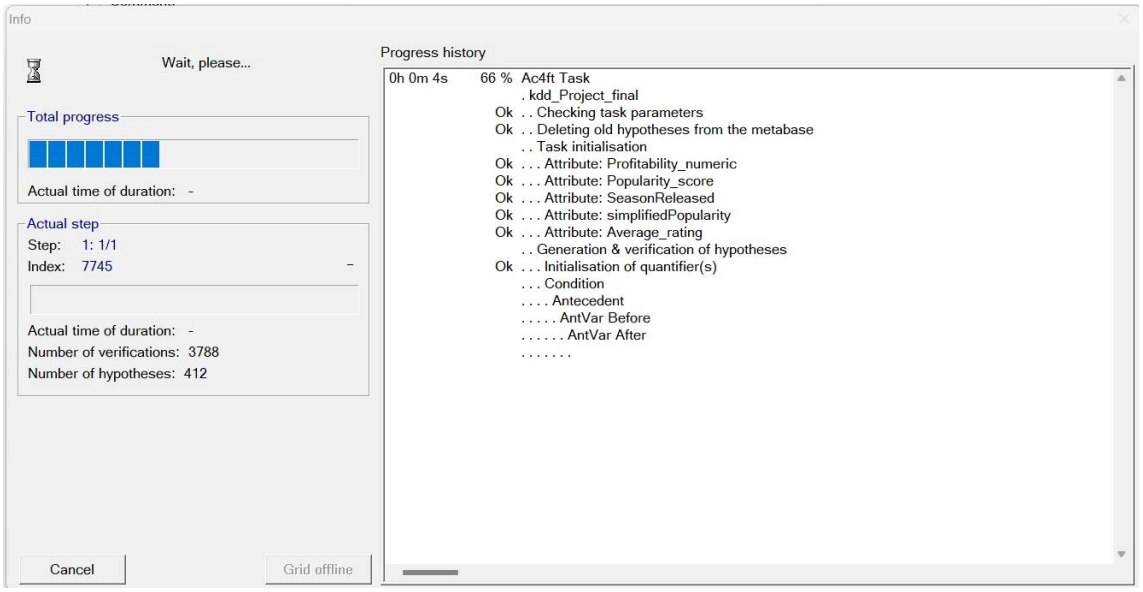


Hence, we can see the Threshold value as 2 in the Quantifiers and In the Succedent Variable part we have changed the coefficient type to “One Category” ranging from 0.5 to 1.5:

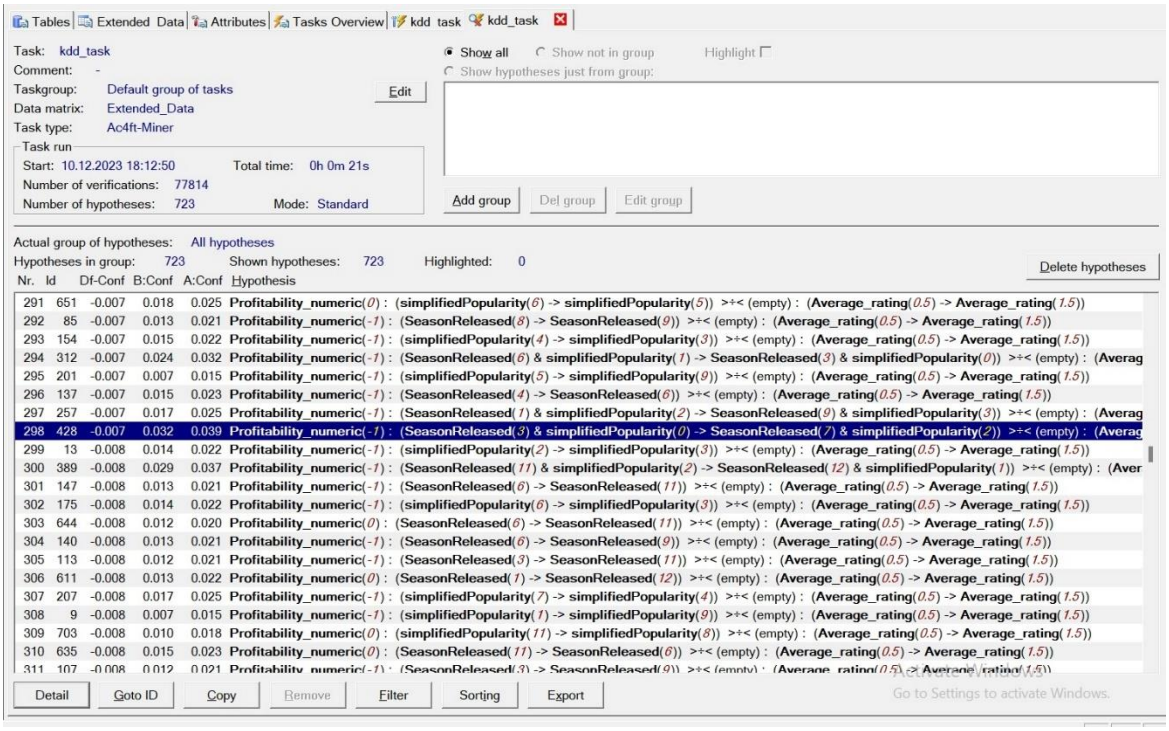
Now, By clicking on “Run” we compute the following:



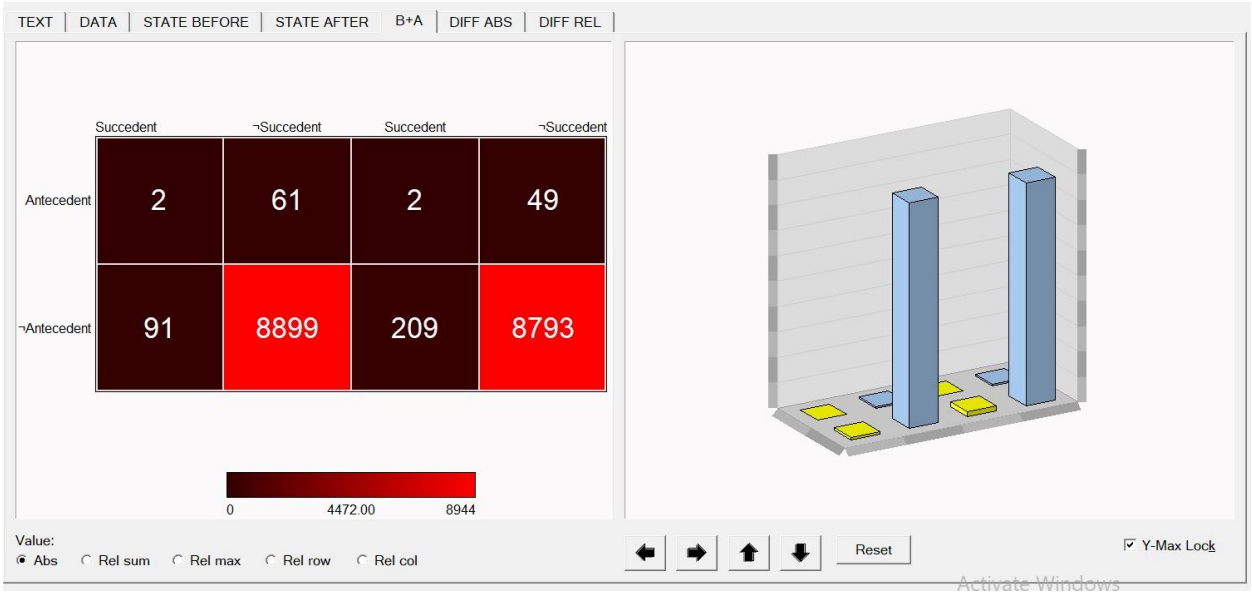
The Task has been initialized and is in progress in the below figure:



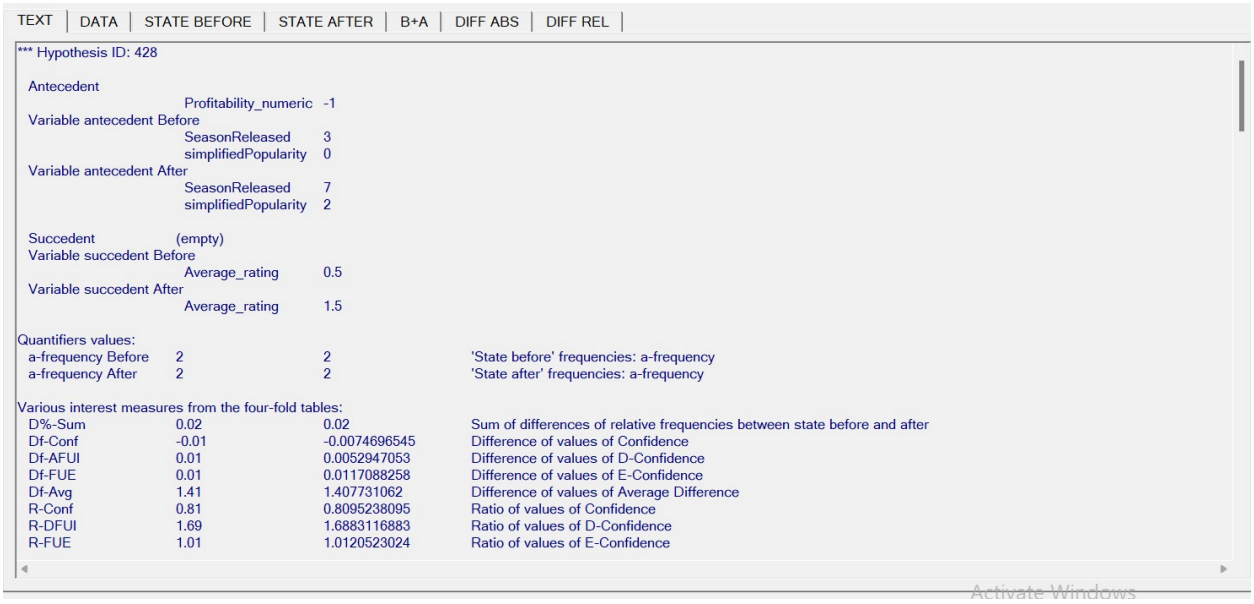
The Results are obtained and stored in a text file as follows:



Lisp Miner - Visual representations of the action rules produced, showcased through screenshots for a detailed and unique perspective.



Text:



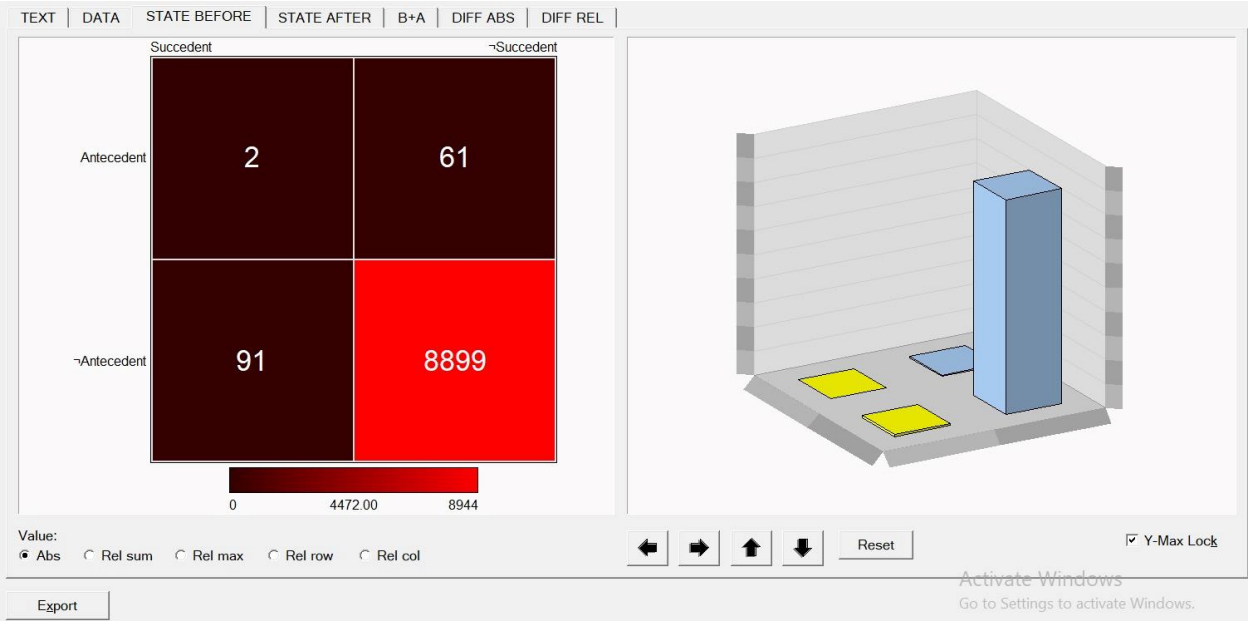
Data:

TEXT | DATA | STATE BEFORE | STATE AFTER | B+A | DIFF ABS | DIFF REL

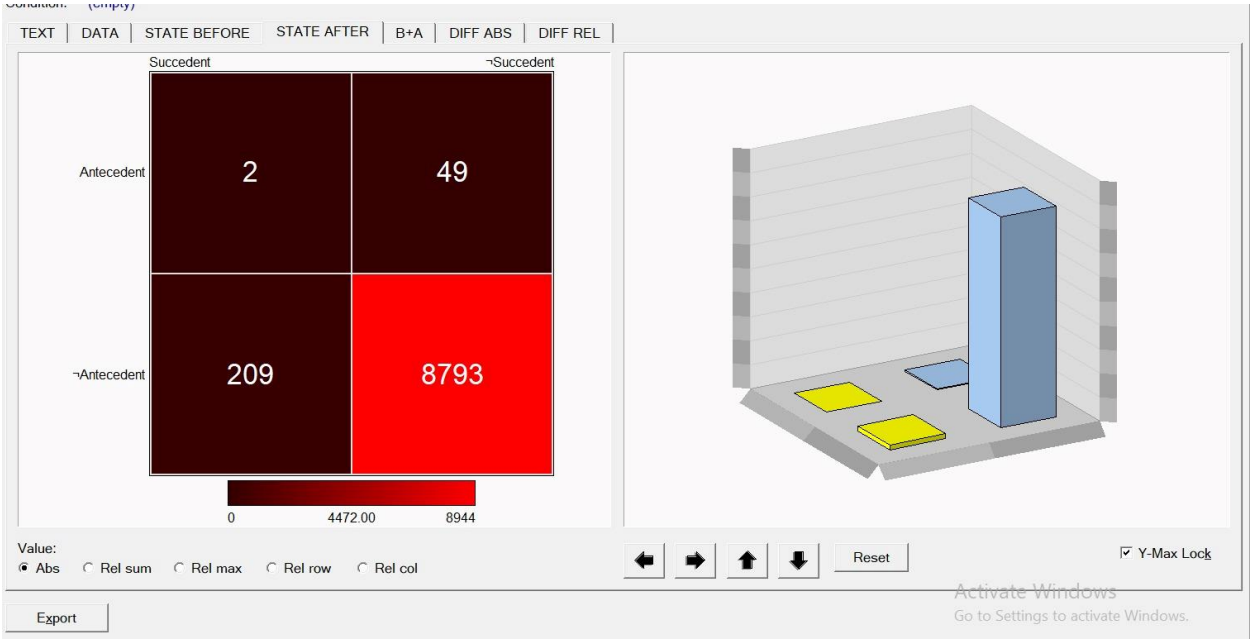
#	A	b	a	S	b	a	Profitability_numeric	SeasonReleased	simplifiedPopularity	SeasonReleased	simplifiedPopularity	Average_rating	Average_rating
1				1			0	10	22	10	22	4	4
2				1			0	12	17	12	17	3.5	3.5
3	1			1			-1	12	12	12	12	3	3
4				1			0	12	4	12	4	2.5	2.5
5				1			0	2	8	2	8	3.5	3.5
6				1			0	12	18	12	18	4	4
7	1			1			-1	12	7	12	7	3.5	3.5
8	1			1			-1	12	3	12	3	4	4
9				1			0	12	5	12	5	3	3
10				1			0	11	15	11	15	3.5	3.5
11				1			0	11	6	11	6	3.5	3.5
12	1			1			-1	12	5	12	5	3	3
13				1			0	12	12	12	12	4	4
14	1			1			-1	12	5	12	5	3.5	3.5
15	1			1			-1	12	7	12	7	2.5	2.5
16				1			0	11	10	11	10	4	4
17				1			0	12	11	12	11	4	4
18				1			0	12	9	12	9	3.5	3.5
19				1			0	11	8	11	8	2.5	2.5
20	1			1			-1	11	7	11	7	2.5	2.5
21				1			0	10	13	10	13	3.5	3.5
22	1			1			-1	10	11	10	11	3.5	3.5
23	1			1			-1	10	11	10	11	3	3
24	1			1			-1	10	12	10	12	3	3
25				1			0	10	10	10	10	3.5	3.5
26	1			1			-1	12	2	12	2	4	4
27				1			0	10	9	10	9	3	3
28	1			1			-1	9	2	9	2	4	4

ACTION RULES:

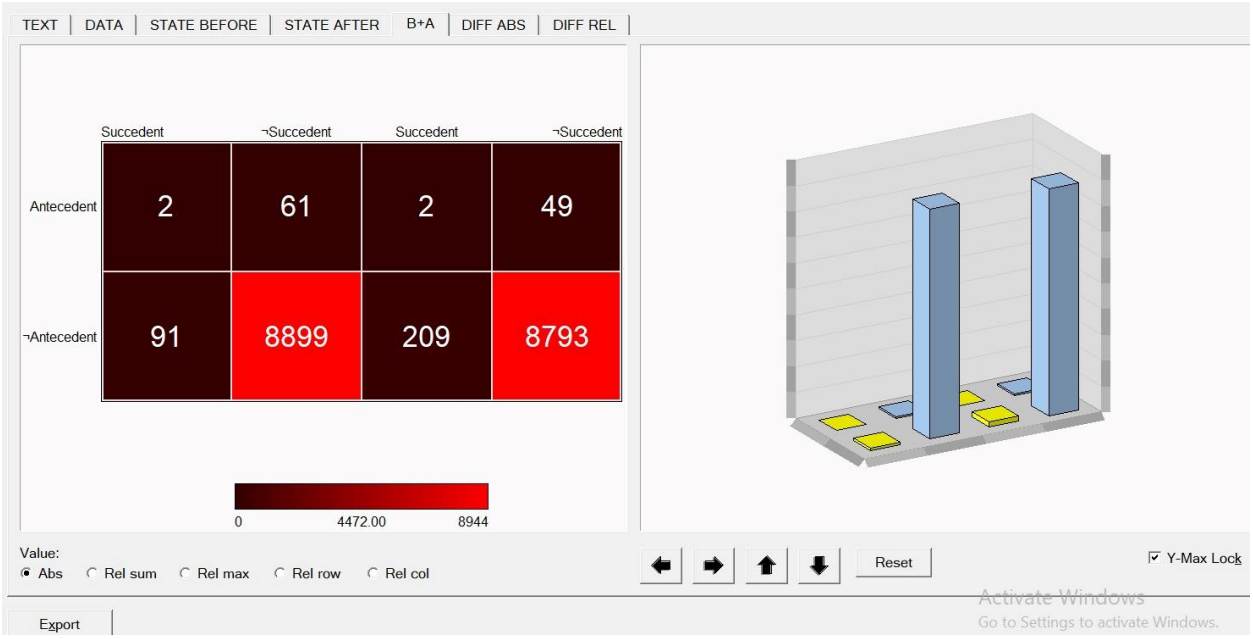
STATE BEFORE:



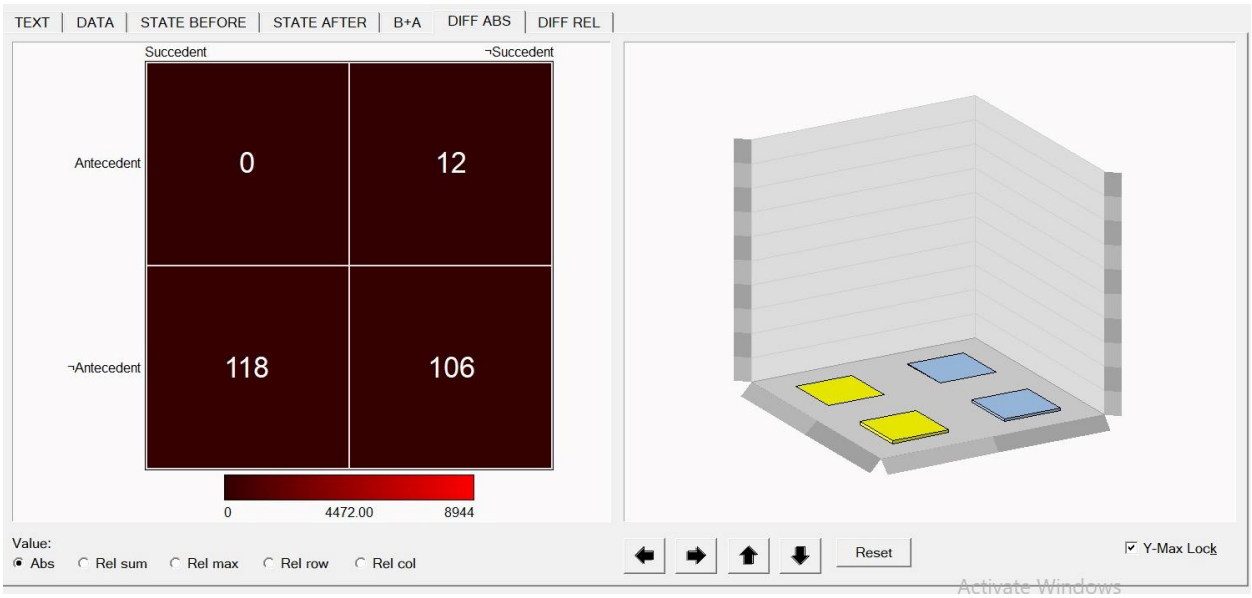
STATE AFTER:



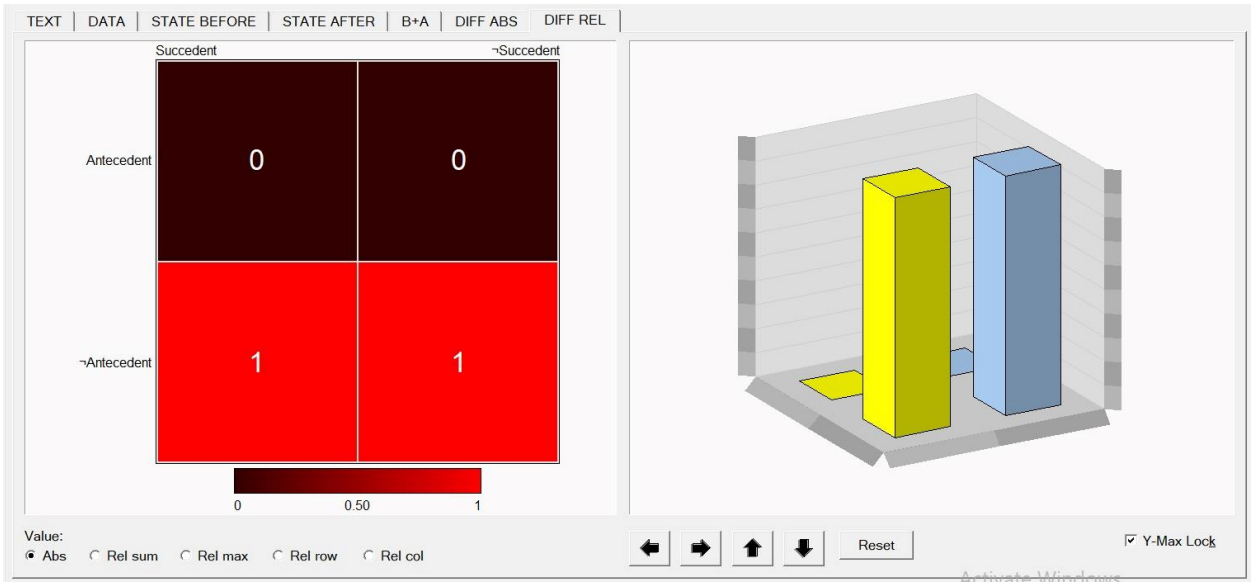
B+A:



DIFF ABS:



DIFF REL:



ANALYSIS OF LISP MINER OUTPUT:

The following significant insights into the factors impacting the average_rating of movies may be obtained from the examination of the 428th hypothesis output from Lisp Miner:

1. **Significant Predictor - Profitability:** "Profitability_numeric" is the most notable predictor, having a significant impact on average_rating. Lower average ratings are linked to movies having a negative profitability_numeric value, which indicates possible losses. This is consistent with the widespread belief that box office performance and viewer happiness are related.

2. **Moderate Predictor - Season Released and Popularity:** The characteristics "simplifiedPopularity" and "SeasonReleased" exhibit a minor influence on average_rating. Seasons in which films are released and popularity levels at which they are released show significant variations in average ratings. This implies that audience reaction is influenced by variables other than financial ones.

3. **Weak Predictor - Simplified Popularity and Profitability:** The characteristics "simplifiedPopularity" and "SeasonReleased" exhibit a minor influence on average_rating. Seasons in which films are released and popularity levels at which they are released show significant variations in average ratings. This implies that audience reaction is influenced by variables other than financial ones.

4. **Least Significant Predictor - Approval:** Initially, "approval" is thought to be the least important predictor. It's possible that films with greater approval ratings don't always convert into better average ratings. This research casts doubt on the widely held notion that audience ratings and favorable reviews are directly correlated.

In summary, the examination indicates that the profitability, release season, and popularity contribute differently to the formation of movie average_ratings. Filmmakers and industry participants can leverage these findings to make well-informed choices regarding movie production, release schedules, and promotional approaches.

CONCLUSION:

Integrating additional components into an established movie dataset, we used WEKA for discretization, preprocessing, and Classification Rules classification. Lisp-Miner was also used to generate Action rules, which serve as enlightening recommendations for raising a film's average rating. These guidelines provide insightful recommendations to improve the overall quality of the movies in the dataset. They were carefully developed, accounting for a number of criteria.

REFERENCES:

- https://waikato.github.io/weka-wiki/downloading_weka/
- https://lispminer.github.io/download/index.html
- https://colab.google/
- https://webpages.uncc.edu/ras/Paper-AR.pdf
- https://lispminer.vse.cz/
- https://www.kaggle.com/datasets/rounakbanik/the-movies-dataset