CS 553
INTELLIGENT DATA ANALYSIS

PROJECT WORKSHOP

ORHUN ALP ORAL

PROJECT OUTLINE

1. **Domain Information**
2. **Dataset: Extraction, Features and possible values**
3. **Preprocessing: Statistics, missing values, outliers**
4. **Analysing Data**
   - To make pattern analysis:
     - Segmentation, clustering, PCA
   - To find explanations:
     - Classification, regression, deviation analysis.
5. **Result Interpretation / Evaluation**
6. **Future Work**

1. DOMAIN INFORMATION

- **Aim of the project is to analyze and understand which factors are more effective for scientific thinking skills for university students.**

- **To understand these factors 1 survey and 4 different psycholohical tests are done to 409 students.**

1. DOMAIN INFORMATION

- **Survey results give demographical information about the students, whereas other tests give scientific results about thinking skills of the students.**

- **The information gathered from analysis are evaluated by Assoc. Prof. Dr. Günseli Orhon from Akdeniz University Institute of Social Sciences Department. She is the expert of this domain and owns the dataset.**

2. DATASET : DATA EXTRACTION

Attitude test towards school:

- Okulda alınan eğitim hayattaki başarıyı garantiler.
- Okulda alınan eğitimle gerçek yaşamdaki başarı ilişkili değildir.
- Çevremde mesleklerinde başarılı olan kişilerin iyi okul deneyimleri var.
- Bence okul zaman kaybıdır. (Negative sentence)
- Okulun en önemli yatırım olduğunu düşünüyorum.
- Okulda alınan eğitim bir insanın kaderini belirler.
- Okula gitmekten genelde keyif alırım.
- Okula gitmekten hoşlanmıyorum.
- Ailem her zaman iyi eğitimin önemli olduğunu düşünür.
- Şu anda okulu bıraksam hayatımda önemli bir değişiklik olmaz.
- Bence okul her insan için gereklidir.
- Okul kazandırması gereken özellikleri kazandıramıyor.
- ...

Scientific thinking test:

- Olaylara bakarken her zaman bilimsel düşünmeye çalışırım
- Bence bilim evrenin sırlarını açıklar
- Bence bilimle çözülemeyecek sorun yok
- Bence bir sorunu çözmenin anahtarı analiz etmektir.
- ...

Academic support:

- Okulda hocalarım bana destek veriyor
- Okulda gerek duyduğum desteği görüyorum
- Okulda sorun yaşadığım zaman hocalarım bunu fark ediyor
- Okulda akademik danışmanımdan memnunum.
- Okuldaki akademik yaşam benim için hayal kırıklığı (Negative meaning)
- Gerek duyduğum her zaman, hocalarıma ulaşabilirim.
- ...

Academic satisfaction:

- Okulda istediğim bilimsel ortamı bulamıyorum
- Okulum tam da hayal ettiğim gibi bir bilimsel altyapı sunuyor
- Okulda iyi öğrendiğimi hissediyorum
- Okulum benim öğrenme ihtiyaçlarımı umursamıyor
- Okulda mutsuzum
- ...

Negative meaning scores are inreversed for evaluation:

- ***Inversion done :***
  - ***1 → 5 , 2 → 4***
  - ***5 → 1 , 4 → 2***

2. DATASET : ATTRIBUTES AND POSSIBLE VALUES

- **There are 408 rows and all 71 features are integers:**
  - Nominal Features (Demographic data from survey results):

  - SED : Socio-Economical Level of the subject. (1..3)
  - GENDER : Gender. (MALE / FEMALE)
  - M_EDU : Educational status of subject's mother. (1..5)
  - F_EDU : Educational status of subject's father. (1..5)

2. DATASET : ATTRIBUTES AND POSSIBLE VALUES

- **There are 408 rows and all 71 features are integers:**

  - AT1..12 : Attitude test towards school. (1..5)
  - AK1..30 : Academic support test. (1..5)
  - AS1..5 : Academic satisfaction scale. (1..5)
  - SC1..20 : Scientific thinking skills test. (1..5)

2. DATASET : DATA CONTENT

Occurances table:

Row ID	S GENDER	I GEND...	S SED	I SED_...	S M_EDU	I M_ED...	S F_EDU	I F_EDU...
Row0	FEMALE	207	1	147	3	130	3	179
Row1	MALE	201	3	143	4	101	4	124
Row2	?	?	2	118	2	99	5	54
Row3	?	?	?	?	1	40	2	51
Row4	?	?	?	?	5	38	?	?
Row5	?	?	?	?	?	?	?	?

Allmost uniformly distributed

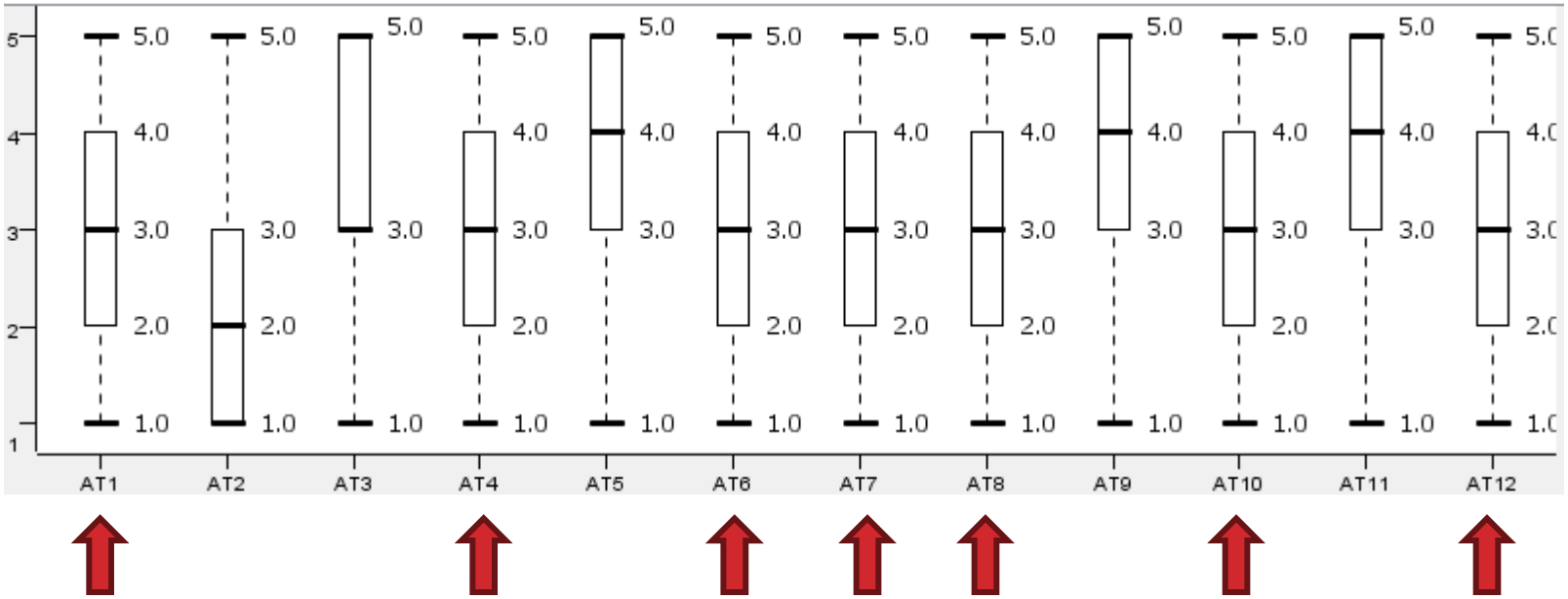
None of the father education status is 1

I AT1	I AT1_...	I AT2	I AT2_...	I AT3	I AT3_...	I AT4	I AT4_...	I AT5	I AT5_...
3	132	1	152	3	117	3	97	5	113
4	115	2	112	5	107	2	95	4	106
2	57	3	99	4	93	1	89	3	104
5	52	4	38	2	50	4	72	2	47
1	52	5	7	1	41	5	54	1	38

Test results' occurancies looks normal..

2. DATASET :
DATA CONTENT BOX PLOT

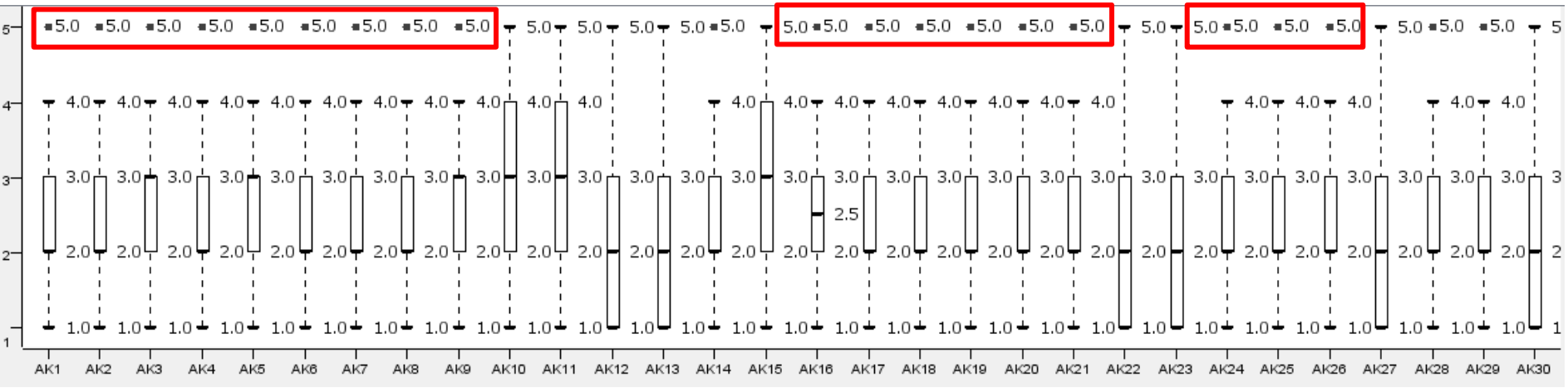
Attitude test towards school AT1..AT12



7 out of 12 questions looks similar distribution.

2. DATASET :
DATA CONTENT BOX PLOT 2

Academic support test AK1..AK30



For most of the questions 5 points are outliers.

Distributions of the test results looks similar.

3. PREPROCESSING: STATISTICS, MISSING VALUES

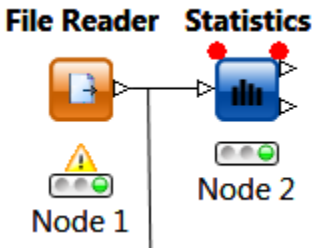
- **Analyse maximum, minimum and missing values.**

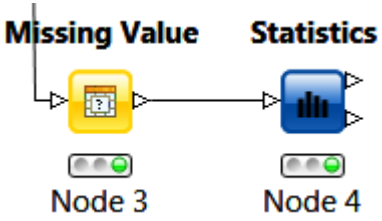
Row ID	SED	M_EDU	F_EDU	AT1	AT2
Minimum	1	1	2	1	1
Maximum	3	5	5	5	5
Mean	1.99	2.995	3.444	3.142	2.108
Std. deviation	0.844	1.122	0.874	1.194	1.065
Variance	0.712	1.258	0.763	1.424	1.133
Overall sum	812	1,222	1,405	1,282	860
No. missings	0	0	0	0	0
Median	?	?	?	?	?
Row count	408	408	408	408	408

with missing values

Row ID	SED	M_EDU	F_EDU	AT1	AT2
Minimum	1	1	2	1	1
Maximum	3	5	5	5	5
Mean	1.992	3	3.449	3.134	2.119
Std. deviation	0.843	1.13	0.869	1.191	1.069
Variance	0.711	1.276	0.754	1.417	1.143
Overall sum	789	1,188	1,366	1,241	839
No. missings	0	0	0	0	0
Median	?	?	?	?	?
Row count	396	396	396	396	396

without missing values



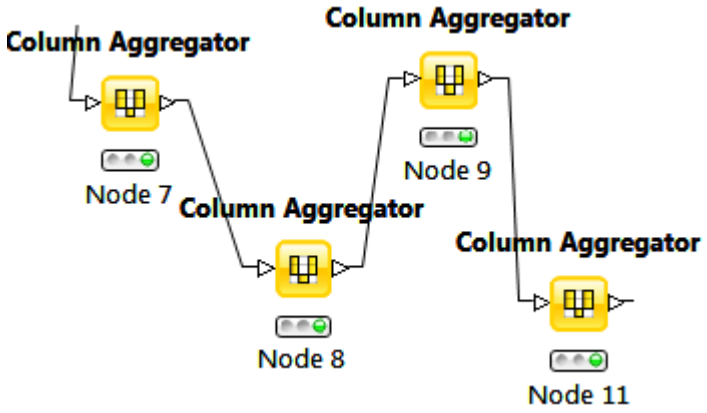


- **%3 of the data is eliminated for further analysis, eliminated data contains missing feature values.**

KNIME toolkit used for this analysis.

3. PREPROCESSING: NEW FEATURE GENERATION

- **For each test, I generate new aggregate feature that shows the sum of that test result.**
- **Because each test feature is actually part of a test group.**
- **SC_SUM feature data contains sum of the SC features values, and it can be the most determinant feature for scientific thinking.**



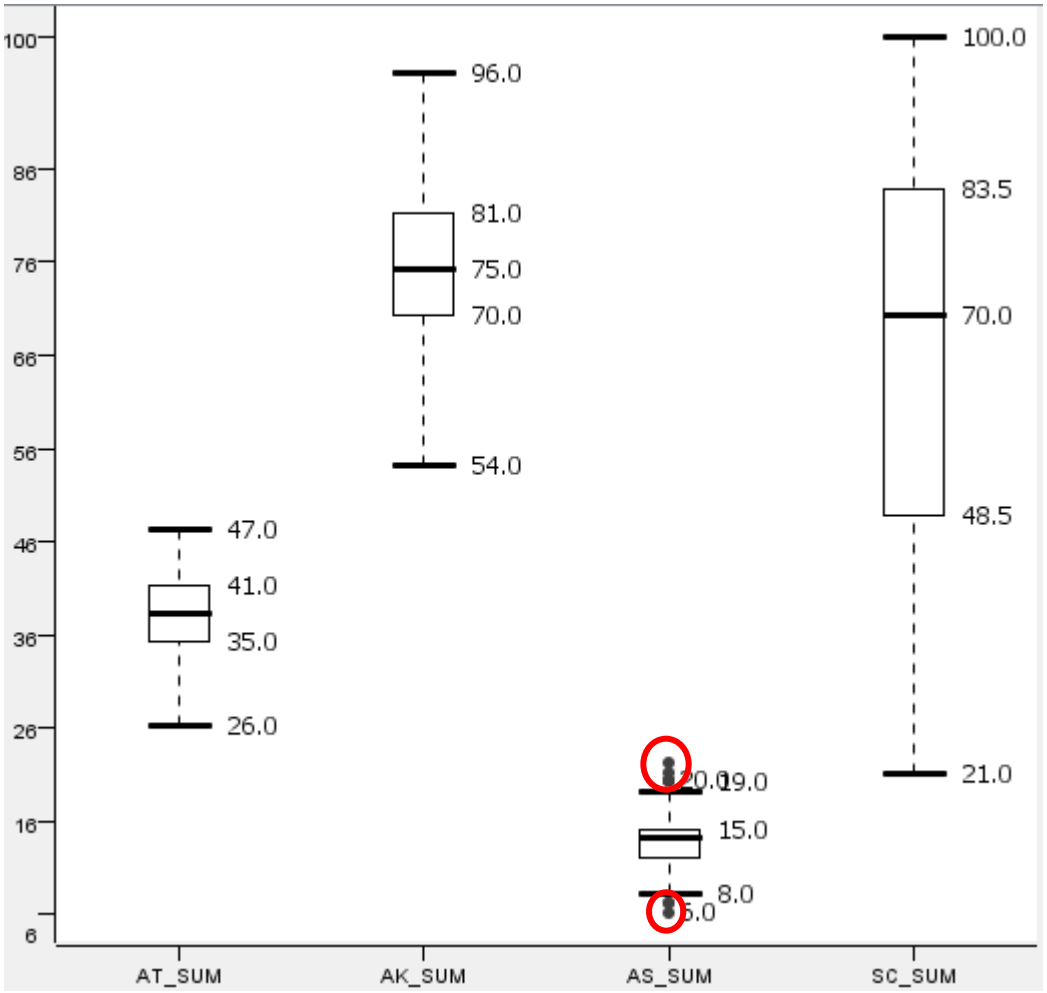
New features are:

- ***AT_SUM***
- ***SC_SUM***
- ***AK_SUM***
- ***AS_SUM***

KNIME toolkit used for this analysis.

3. PREPROCESSING: OUTLIER VERIFICATION

- **Aggregated test results analyzed in the box plot.**
- **Only academic support test has some outliers.**
- **These outliers are examined and verified their truth existence by the expert.**

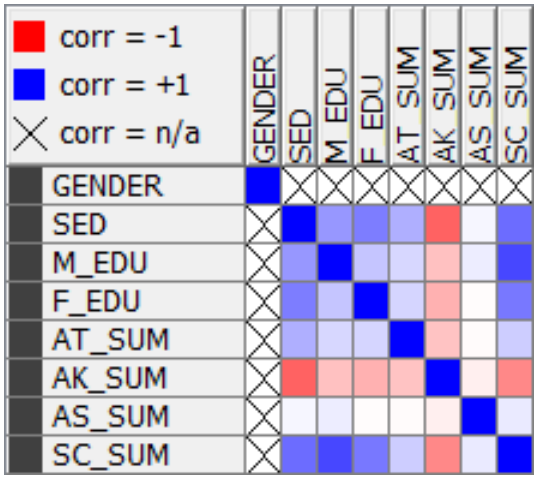


KNIME toolkit used for this analysis.

4. ANALYSING DATA: CORRELATION FOR GENERATED FEATURES

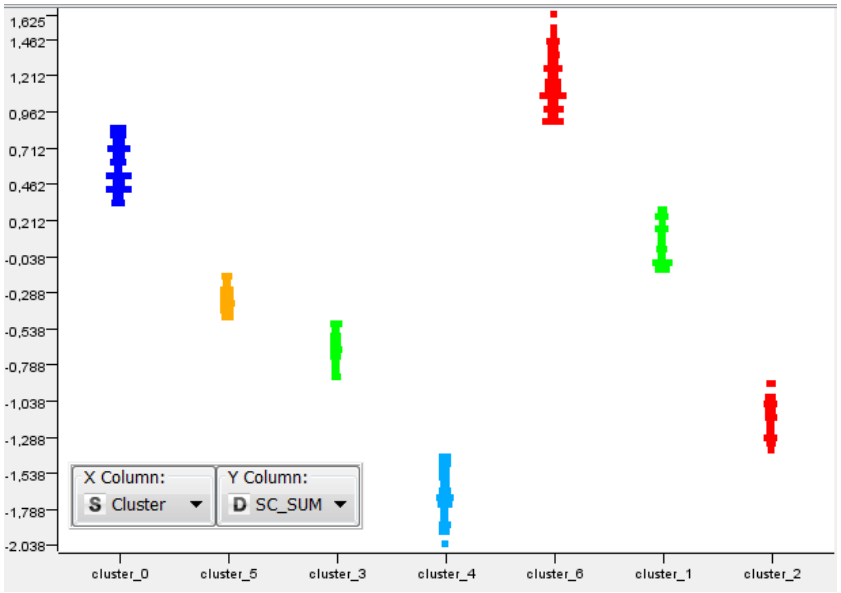
There are some significant correlations between features generated features and demographic information:

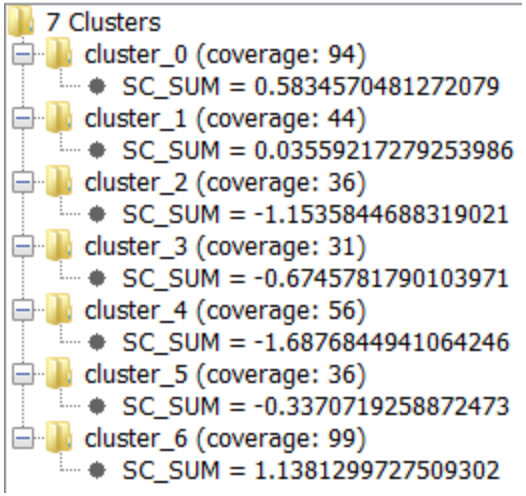
- **Socio-eco. status is negatively correlated with academic support test.**
- **Mother's education status is positively correlated with scientific thinking test.**



4. ANALYSING DATA: TO MAKE PATTERN ANALYSIS

- **To detect groups in the different scientific thinking test, k-means algorithm used with k=7 with normalized SC_SUM**

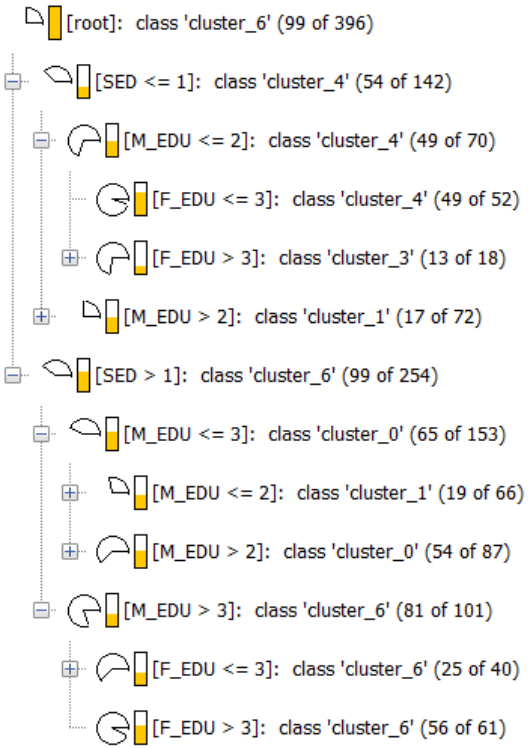




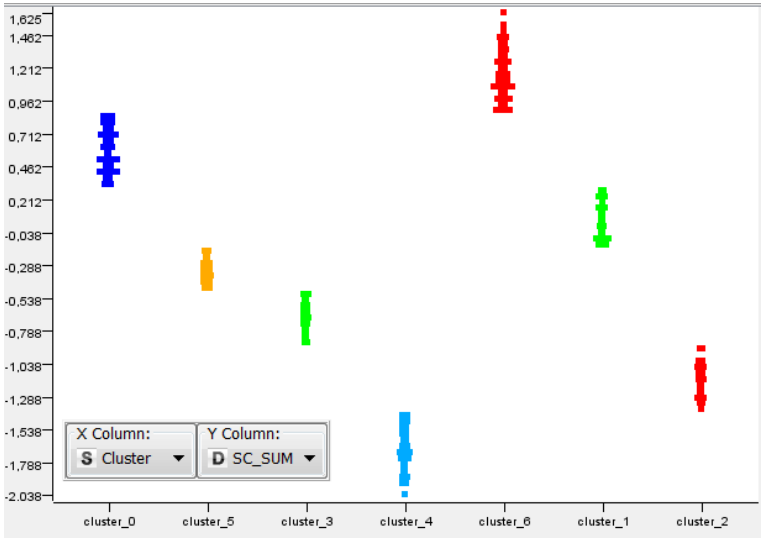
KNIME toolkit used for this analysis.

4. ANALYSING DATA: TO MAKE PATTERN ANALYSIS

- **To analyse the relation of the clusters with respect to demographical information about the student, use decision tree**
  - *Socio-economic status, gender, father education, mother education.*

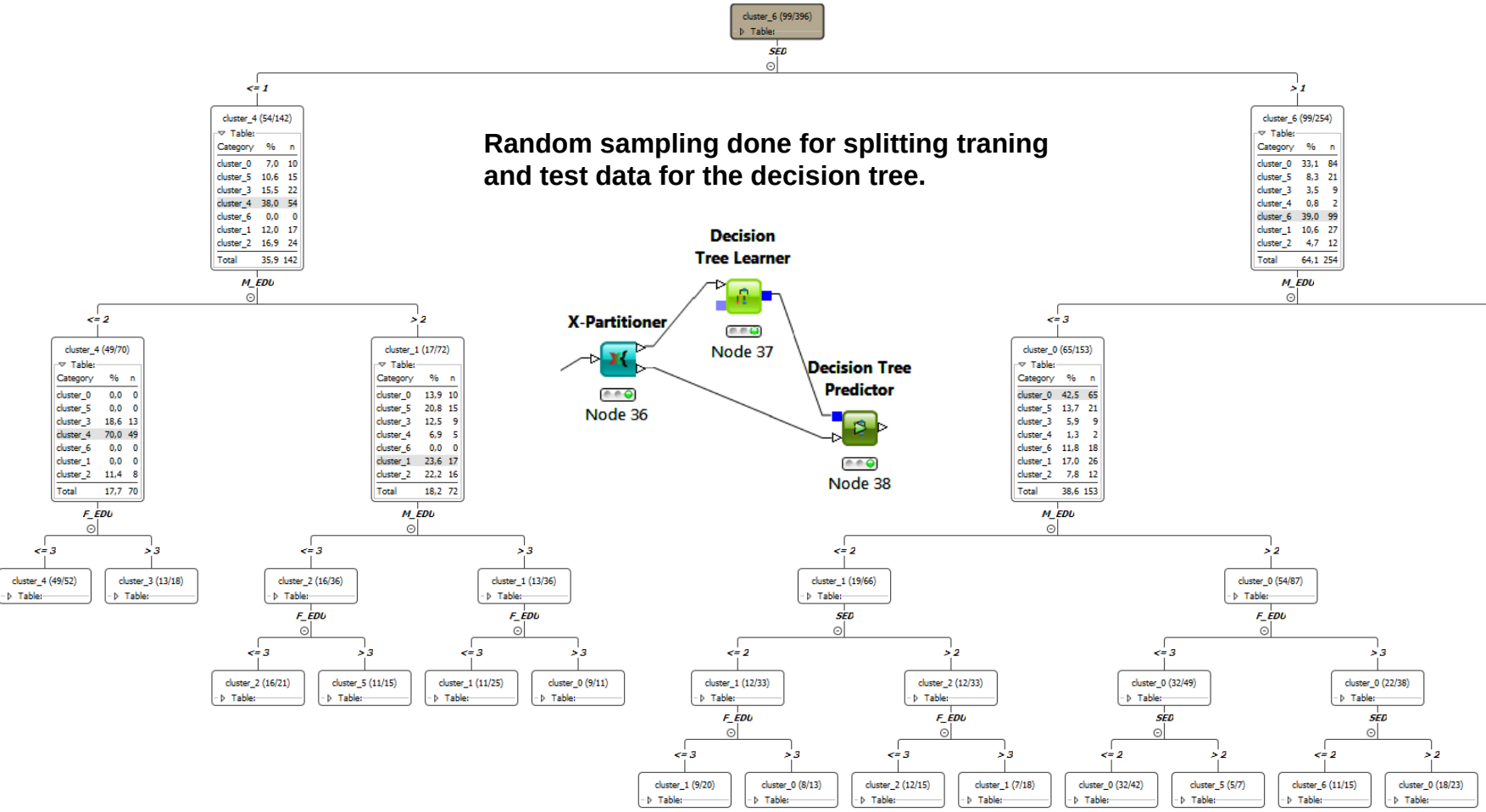


Decision tree shows that socio-economic status is the significant factor between separation of ***cluster_6*** *from* ***cluster_4****.*



KNIME toolkit used for this analysis.

4. ANALYSING DATA: TREE VIEW C4.5



Random sampling done for splitting traning and test data for the decision tree.

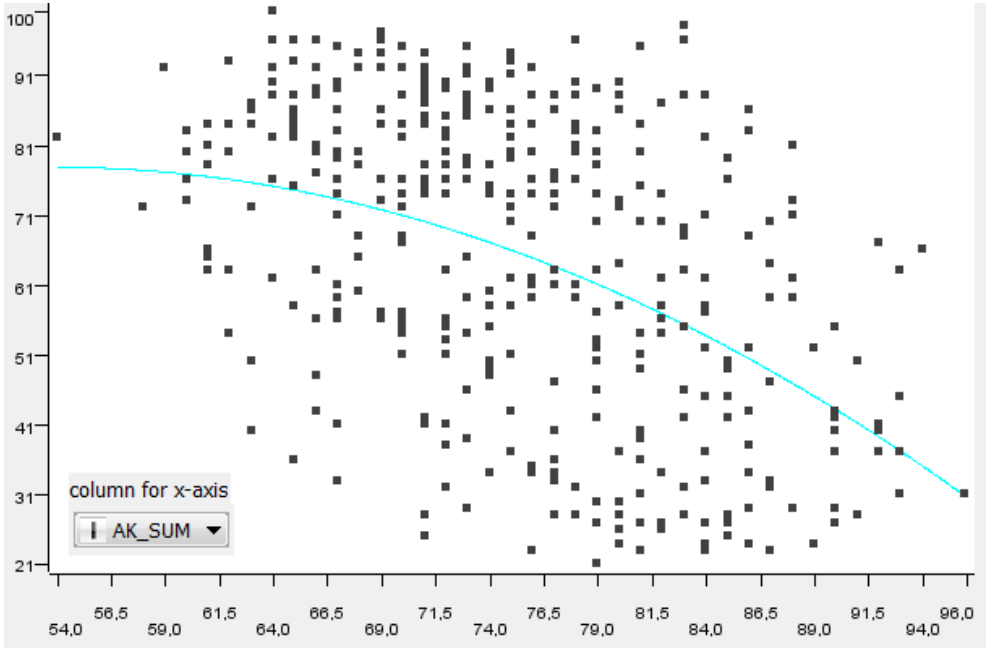
4 . ANALYSING DATA:
TO FIND EXPLANATIONS (REGRESSION)

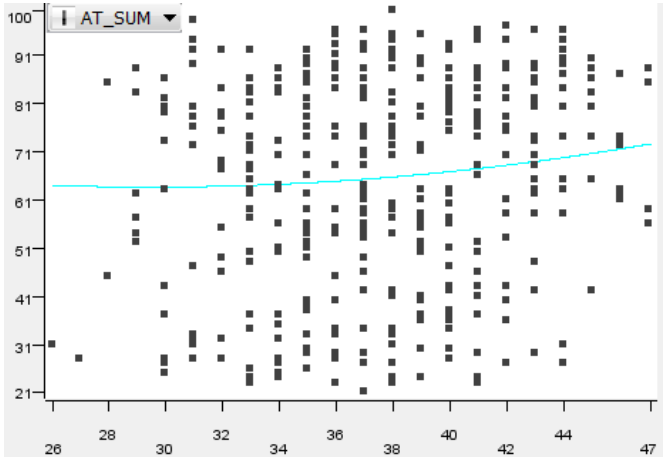
Regression with generated columns

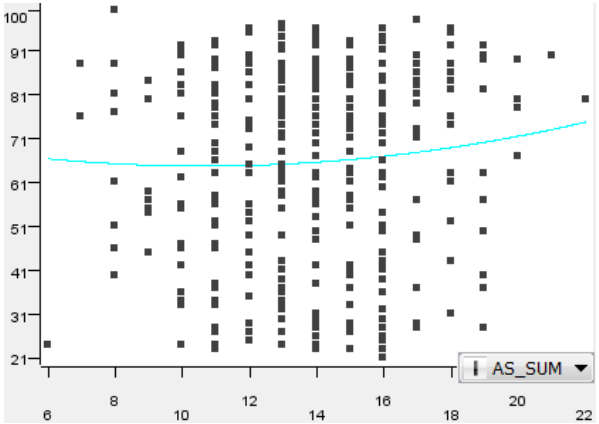
Target: SC_SUM

Attribute name	Intercept	Coefficient (x^1)	Coefficient (x^2)
AT_SUM	32,4830	-1,7194	0,0292
AK_SUM	0,0292	2,8546	-0,0265
AS_SUM	-0,0265	-1,6439	0,0774

Squared error (per row): 352.75147188341185







Most significant characteristics is for AK_SUM

4 . ANALYSING DATA:
TO FIND EXPLANATIONS (REGRESSION)

Regression with generated columns

Target: SC_SUM

Parameters: AT_SUM, AK_SUM, AS_SUM

Columns: 4	PolyReg prediction	...	Prediction Error	...
Column Type	DoubleCell	I...	DoubleCell	I...
Column Index	0	1	2	3
Color Handler				
Size Handler				
Shape Han...				
Lower Bound	29.290698446242...	1	0.13287491331...	1
Upper Bound	85.21331292775535	2	45.2297888645...	1

Row ID	D PolyReg prediction	D Prediction Error
Minimum	29.291	0.133
Maximum	85.213	45.23
Mean	64.937	15.74
Std. deviation	10.562	10.261
Variance	111.564	105.28
Overall sum	25,715	6,232.908
No. missings	0	0
Median	?	?
Row count	396	396

Min error rate : 0.13 ➔~%0
Max error rate : 45.22 ➔%57

Mean prediction error: 15.74
Scale for SC_SUM: 20-100
Mean error rate: %19

4 . ANALYSING DATA:
TO FIND EXPLANATIONS (REGRESSION)

Regression with existing columns

Target: SC_SUM

Parameters:

- AT1..12
- AK1..30
- AS1..5

Attribute name	Intercept	Coefficient (x^1)	Coefficient (x^2)
AT1	88,6873	2,6022	-0,1629
AT2	-0,1629	-6,2305	0,8415
AT3	0,8415	2,1441	-0,2005
AT4	-0,2005	-1,2109	0,0899
AT5	0,0899	-8,2938	1,3083
AT6	1,3083	13,4994	-2,2424
AT7	-2,2424	2,1564	-0,2171
AT8	-0,2171	4,9615	-0,9982
AT9	-0,9982	-0,0154	0,1065
AT10	0,1065	-1,0034	0,1386
AT11	0,1386	5,4564	-0,6965
AT12	-0,6965	-0,9481	0,2537
AK1	0,2537	-8,5459	0,9454
AK2	0,9454	3,2884	-1,1580
AK3	-1,1580	0,3543	0,0025
AK4	0,0025	10,7285	-2,2660
AK5	-2,2660	0,9559	-0,3620
AK6	-0,3620	2,4859	-0,3260
AK7	-0,3260	4,8746	-1,4741
AK8	-1,4741	0,9960	-0,3908
AK9	-0,3908	-2,0805	0,3058
AK10	0,3058	2,6400	-0,1148
AK11	-0,1148	2,1947	-0,4262
AK12	-0,4262	5,9566	-1,4038
AK13	-1,4038	-11,7183	2,1619
AK14	2,1619	-4,9618	0,9000
AK15	0,9000	-2,8113	0,2250
AK16	0,2250	5,7215	-1,1767
AK17	-1,1767	-5,2772	0,8285
AK18	0,8285	2,8924	-0,8437
AK19	-0,8437	-6,8018	0,7866
AK20	0,7866	-4,7319	0,2796
AK21	0,2796	-0,2899	-0,0397
AK22	-0,0397	3,2348	-0,8284
AK23	-0,8284	-13,8386	2,6751
AK24	2,6751	-0,2758	0,0618
AK25	0,0618	-7,4323	0,9771
AK26	0,9771	-2,6740	0,7945
AK27	0,7945	-7,8163	1,4621
AK28	1,4621	0,7311	-0,4835
AK29	-0,4835	2,8562	-0,8138
AK30	-0,8138	-0,1183	0,1979
AS1	0,1979	7,9401	-1,0102
AS2	-1,0102	4,1604	-0,5743
AS3	-0,5743	1,0724	-0,2794
AS4	-0,2794	0,0184	0,2332
AS5	0,2332	-1,1493	0,2218

Squared error (per row): 241.45892147322866

4 . ANALYSING DATA:
TO FIND EXPLANATIONS (REGRESSION)

Regression with generated columns

Target:SC_SUM

Parameters: all test parameters: AK1..30, AT1.12, AS1..5

Columns: 4	PolyReg prediction	Pol...	Prediction Error	Pre...
Column Type	DoubleCell	IntCell	DoubleCell	Int...
Column Index	0	1	2	3
Color Handler				
Size Handler				
Shape Han...				
Lower Bound	25.658940483203...	1	0.02570019726...	1
Upper Bound	97.515652791381	1	50.4427077343...	1

Upper bound increased by 12 compare to previous case
Lower bound decreased by 4 compare to previous case

Min error rate : ~0
Max error rate : 50.44 ➔%63.05

Row ID	D PolyReg prediction	D Prediction Error
Minimum	25.659	0.026
Maximum	97.516	50.443
Mean	54.937	12.769
Std. deviation	14.938	8.866
Variance	223.138	78.607
Overall sum	25,715	5,056.57
No. missings	0	0
Median	?	?
Row count	396	396

Mean prediction error: 12.43
Scale for SC_SUM: 20-100
Mean error rate: %15

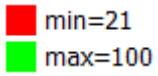
%4 decreased

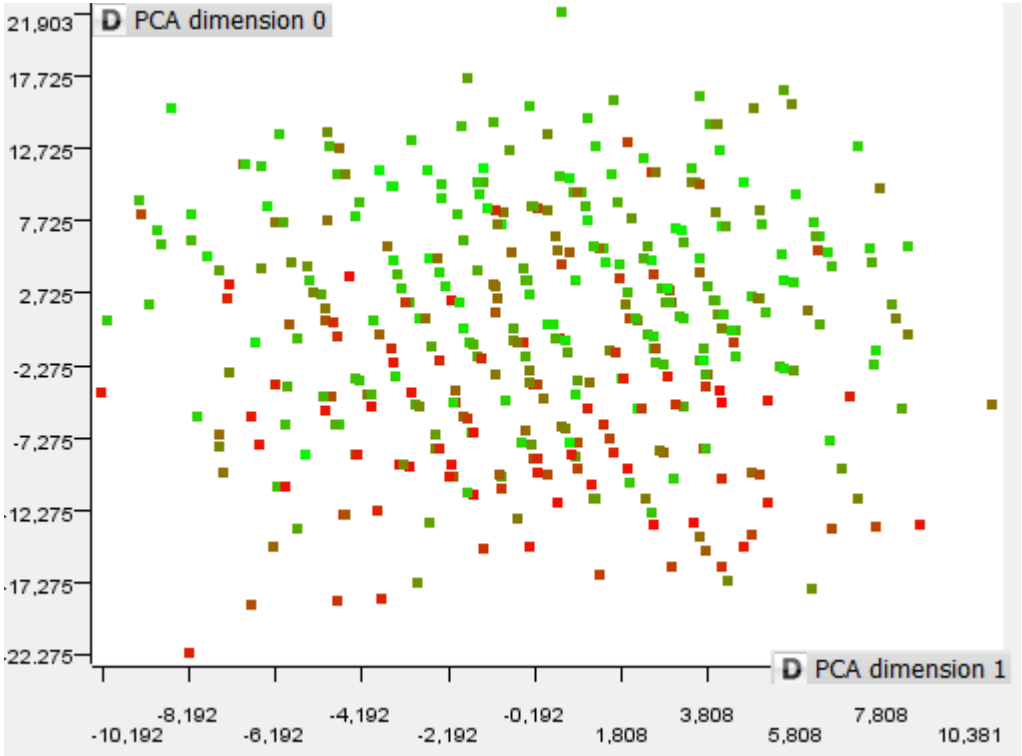
4 . ANALYSING DATA: TO FIND PATTERN (PCA)

PCA analysis with:

AT_SUM, AK_SUM, AS_SUM

For SC_SUM:





To search that whether is it possible to find pattern of scientific thinking test result from variability between other test results.

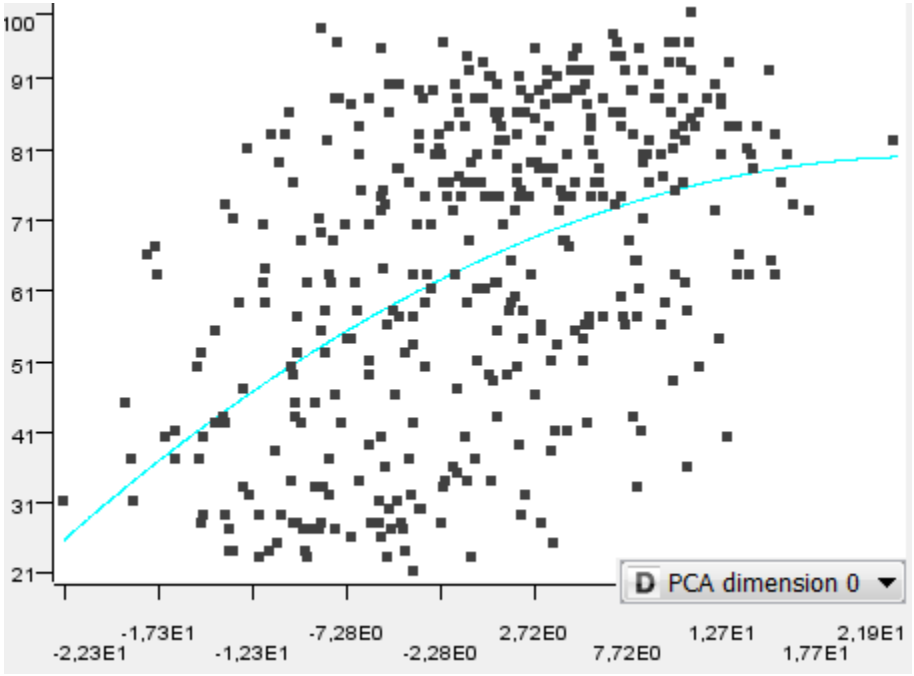
4 . ANALYSING DATA:
TO FIND PATTERN (PCA-REGRESSION)

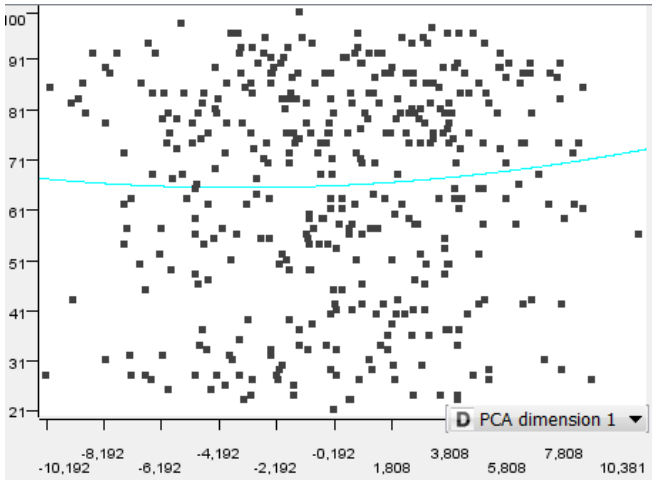
Regression with pricipal components

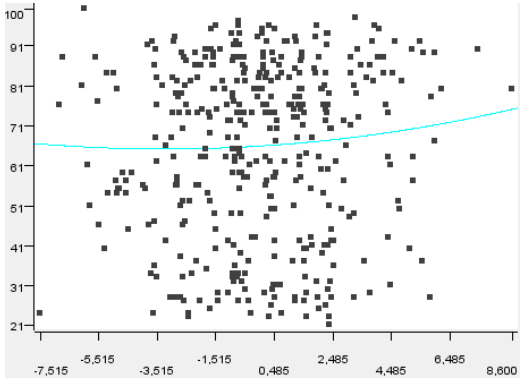
Target: SC_SUM

Attribute name	Intercept	Coefficient (x^1)	Coefficient (x^2)
PCA dimension 0	65,4769	1,2164	-0,0254
PCA dimension 1	-0,0254	0,2642	0,0389
PCA dimension 2	0,0389	0,4528	0,0713

Squared error (per row): 352.5277757715358







Most significant characteristics is for PCA_dimension 0

4 . ANALYSING DATA:
TO FIND PATTERN (PCA-REGRESSION)

Regression with principle components

Target: SC_SUM

Parameters: PCA_dim0, PCA_dim1, PCA_dim2

Columns: 4	PolyReg prediction	PolyRe...	Prediction Error	Predi...
Column Type	DoubleCell	IntCell	DoubleCell	IntCell
Column Index	0	1	2	3
Color Handler				
Size Handler				
Shape Han...				
Lower Bound	26.42150636065...	1	0.0571247715...	1
Upper Bound	83.95532319180...	2	45.502900929...	1

Row ID	D PolyReg prediction	D Prediction Error
Minimum	26.422	0.057
Maximum	83.955	45.503
Mean	64.937	15.726
Std. deviation	10.573	10.27
Variance	111.788	105.474
Overall sum	25,715	6,227.664
No. missings	0	0
Median	?	?
Row count	396	396

Min error rate : ~0
Max error rate : 45.50 ➔%56.05

Mean prediction error: 15.72
Scale for SC_SUM: 20-100
Mean error rate: %19

Same results with test totals' regression

5. RESULT INTERPRETATION / EVALUATION

- **Analysis shows that scientific thinking skills change with the following:**
  - ✓ If the social-economic status of the student is middle level; it increases, whereas low social-economic status has negative effects.
  - ✓ Parent's educational status is positively correlated.
  - ✓ High attitude test results show positive effects on scientific thinking skills.
  - ○ High academic satisfaction skills show positive effects on scientific thinking skills.

6. FUTURE WORK

- **Polynomial combinations can be used for regression analysis with assigning different weights to each feature.**

- **Other test results' reasons should be investigated to find out the effects of the features more detailed.**

QUESTIONS?

Thank you.