Reference

[1] Adnan, M.H.M., Husain, W., Rashid, N.A.A, (2012). Data Mining for Medical Systems: A Review. International Conference on Advances in Computer and Information Technology. , pp.17-22

[2] Ahmadvand, A. M., Bidgoli, B. M., & Akhondzadeh, E. (2010, January). A hybrid data mining model for effective citizen relationship management: a case study on Tehran municipality. In e-Education, e-Business, e-Management, and e-Learning, 2010. IC4E'10. International Conference on (pp. 277-281). IEEE

[3] Asian Development Bank.(2013).Asian Development Bank Annual Report 2013. [ONLINE] Available at: http://www.adb.org/sites/default/files/institutional-document/42741/adb-annual-report-2013.pdf. [Accessed 09 July 15].

[4] Atkinson, B., & Marlier, E. (2010). Income and living conditions in Europe

[5] Ayanso, A., Lertwachara, K., & Vachon, F. (2011). Design and behavioral science research in premier IS journals: evidence from database management research. In Service-oriented perspectives in design science research (pp. 138-152). Springer Berlin Heidelberg

[6] Berkhin, P. (2006). A survey of clustering data mining techniques. In*Grouping multidimensional data* (pp. 25-71). Springer Berlin Heidelberg.

[7] Business Intelligence Solutions.(2015). Data Mining vs. Statistics. [ONLINE] Available at: http://www.bisolutions.us/Data-Mining-vs-Statistics.php. [Accessed 09 July 15].

[8] Cardoso, A. R., & Verner, D. (2006). School drop-out and push-out factors in Brazil: The role of early parenthood, child labor, and poverty.

[9] Central Bank of Sri Lanka.(2013). Central Bank of Sri Lanka Annual Report 2013.[ONLINE] Available at: http://www.cbsl.gov.lk//pics_n_docs/10_pub/_docs/efr/annual_report/ar2013/english/content.htm. [Accessed 09 July 15].

[10] Chung, Y., 2013. Chronic Health Conditions and Economic Outcomes.

[11] Department of Census and Statistics. (2013). Household Income and Expenditure Survey -2012/2013 Final Results. [ONLINE] Available at:

http://www.statistics.gov.lk/HIES/HIES200213FinalBuletin4.pdf. [Accessed 09 July 15].

[12] Ec.europa.eu, (2015). [online] Available at: http://ec.europa.eu/eurostat/documents/3217494/5722557/KS-31-10-555-EN.PDF/e8c0a679-be01-461c-a08b-7eb08a272767 [Accessed 11 Nov. 2015].

[13] Einsele, F., Sadeghi, L., Ingold, R., & Jenzer, H. (2015). A Study about Discovery of Critical Food Consumption Patterns Linked with Lifestyle Diseases using Data Mining Methods

[14] Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P., & Uthurusamy, R. (1996). Advances in knowledge discovery and data mining

[15] Fernandez-Villaverde, J., & Krueger, D. (2007). Consumption over the life cycle: Facts from consumer expenditure survey data. The Review of Economics and Statistics, 89(3), 552-565

[16] Hamel, L., & Hall, T. (2005).A brief tutorial on database queries, data mining, and OLAP. The Encyclopedia of Data Warehousing and Mining, 401.

[17] Hildebrandt, M., & Gutwirth, S. (2008). Profiling the European citizen. Heidelberg: Springer

[18] Jiang, S., Ferreira, J., & González, M. (2013, January). ANALYZING HOUSEHOLD LIFESTYLES, MOBILITY AND ACTIVITY PROFILES: A CASE STUDY OF SINGAPORE. 92nd Annual Meeting

[19] Jung, Y. G., Kang, M. S., & Heo, J. (2014). Clustering performance comparison using K-means and expectation maximization algorithms.*Biotechnology & Biotechnological Equipment*, *28*(sup1), S44-S48.

[20] Junling, H. (2013). About Data Mining: Data Mining vs. Machine Learning. [ONLINE] Available at: http://www.aboutdm.com/2013/02/data-mining-vs-machine-learning.html. [Accessed 09 July 15]

[21] Khan, M. A., Islam, Z., & Hafeez, M. (2011, December). Irrigation water demand forecasting: a data pre-processing and data mining approach based on spatio-temporal data. In Proceedings of the Ninth Australasian Data Mining Conference-Volume 121 (pp. 183-194). Australian Computer Society, Inc.

[22] Koh, H. C.,Tang,G., (2011). Data Mining for Medical Systems: A Review. Journal of Healthcare Information Management. 19 (2), pp.64-72 -19

[23] Lee, H. (2007). Essentials of Behavioral Science Research

[24] Mahrsi, M. K. E., Etienne, C. O. M. E., Johanna, B. A. R. O., & Oukhellou, L. (2014, January). Understanding Passenger Patterns in Public Transit Through Smart Card and Socioeconomic Data: A case study in Rennes, France. In ACM SIGKDD Workshop on Urban Computing (p. 9p)

[25] Mina, C. D., & Barrios, E. B. (2009). Profiling poverty with multivariate adaptive regression splines. Philippine Institute for Development Studies.

[26] Okumu, I. M., Nakajjo, A., & Isoke, D. (2008). Socioeconomic determinants of primary school dropout: the logistic model analysis

[27] Orodho, A. J., & Kombo, D. K. (2002). Research methods. Nairobi: Kenyatta University, Institute of Open Learning

[28] Publications, S. (2015). Home &middot Science Publications. [online] Thescipub.com. Available at: http://thescipub.com/PDF/ajassp.2009.2036.2042.pdf [Accessed 11 Nov. 2015].

[29] Rahman, H. (Ed.). (2008). Data mining applications for empowering knowledge societies. IGI Global.

[30] Rokach, L., & Maimon, O. (2005). Clustering methods. In *Data mining and knowledge discovery handbook* (pp. 321-352). Springer US

[31] Smith, J.P., 1999. Healthy Bodies and Thick Wallets: The Dual Relation Between Health and Economic Status [WWW Document]. URL https://www.aeaweb.org/articles.php?doi=10.1257/jep.13.2.145 (accessed 1.24.16).

[32] UNESCO Institute for Statistics. (2014) GUIDE TO THE ANALYSIS AND USE OF HOUSEHOLD SURVEY AND CENSUS EDUCATION DATA. [ONLINE] Available at: www.uis.unesco.org/Library/Documents/hhsguide04-en.pdf. [Accessed 09 July 15].

[33] UNICEF, (2013). Out of School Children in Sri Lanka, Summary Report. [online] UNICEF Sri Lanka. Available at: http://www.unicef.org/srilanka/2013_OSS_Summery_E.pdf [Accessed 10 Nov. 2015]

[34] Warc.com, (2015). Lifestyle segmentation of the Chinese consumer. [online] Available at: http://www.warc.com/fulltext/esomar/80217.htm [Accessed 11 Nov. 2015].

[35] Wirth, R., &Hipp, J. (2000). CRISP-DM: Towards a standard process model for data mining. In Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining (pp. 29-39)

[36] World Bank Group. (2015). Assessing Sector Performance and Inequality in Education –Chapter 2. [ONLINE] Available at: http://siteresources.worldbank.org/EXTEDSTATS/Resources/3232763-1252439241095/ADePTBookChap-2.pdf [Accessed 09 July 15].

Appendix A

Data Preprocessing

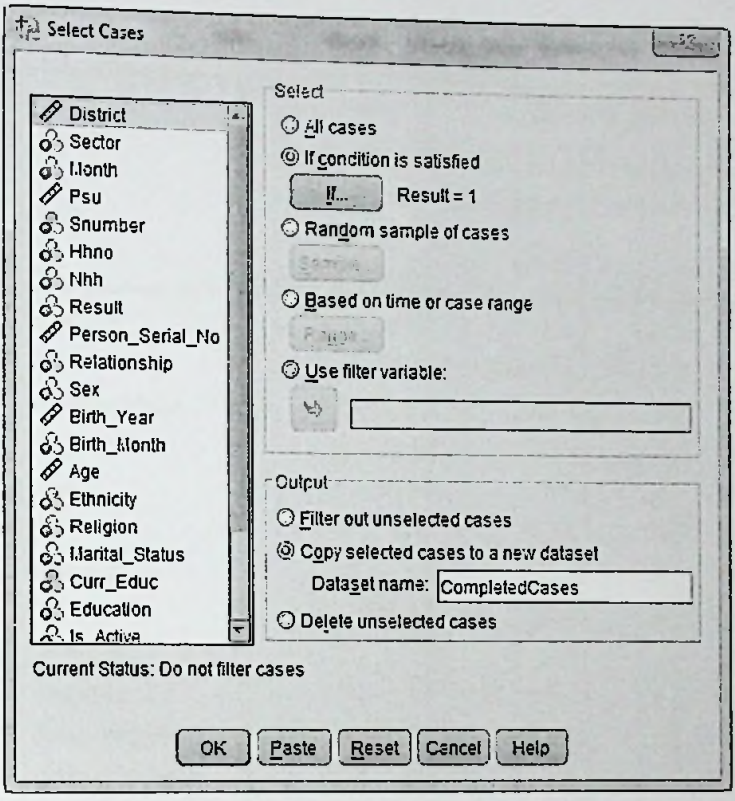

Figure 9.1: ignoring the tuple

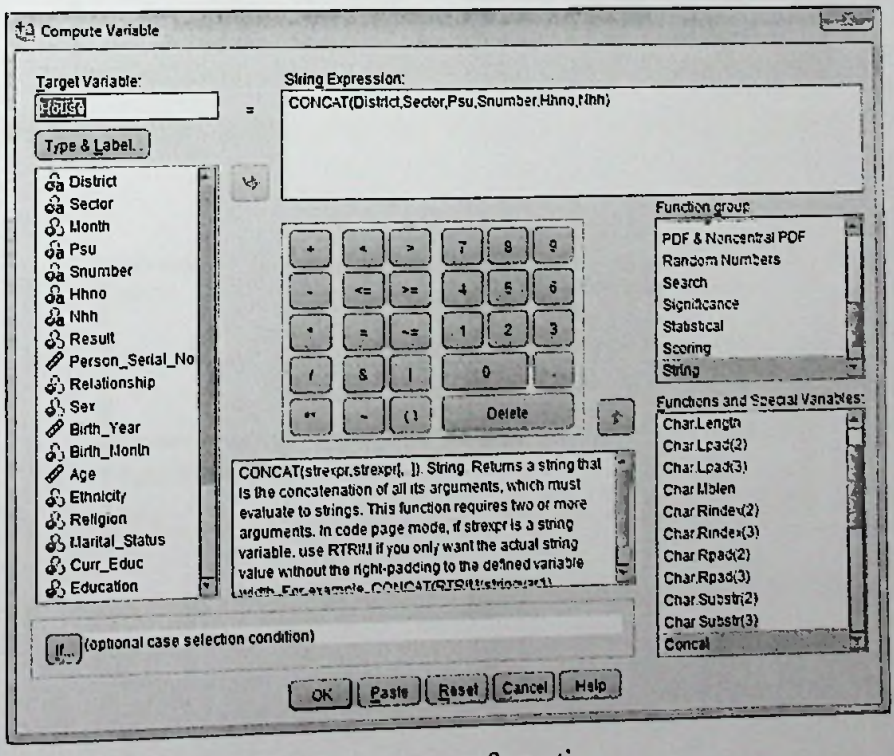

Figure 9.2: data transformation

Appendix B

Attribute Selection using SPSS

Likelihood Ratio Tests

Effect	Model Fitting Criteria	Likelihood Ratio Tests		
	-2 Log Likelihood of Reduced Model	Chi-Square	df	Sig.
Intercept	5776.207[a]	.000	0	.
Age	9245.118[b]	3468.912	2	.000
HouseholdSize	5782.583[b]	6.376	2	.041
District	5889.220[b]	113.013	48	.000
Sector	5778.631[b]	2.425	4	.658
Sex	5780.354[b]	4.147	2	.126
Ethnicity	5784.428[b]	8.221	12	.768
Religion	5789.936[b]	13.730	8	.089
Is_Active	6603.556[b]	827.350	2	.000
HeadSex	5785.470[b]	9.263	2	.010
FatherEducation	5898.866[b]	122.659	42	.000
MotherEducation	5977.752[b]	201.545	42	.000
income	5779.287[b]	3.080	4	.545
Structure	5803.527[b]	27.321	20	.126
Natural_Calamity	5777.901[b]	1.694	2	.429
Is_Ill_Disable11	6036.825[b]	260.618	2	.000

Figure 9.3: Likelihood Ratio Test considering schooling and dropouts both

Model Fitting Information

Model	Model Fitting Criteria	Likelihood Ratio Tests		
	-2 Log Likelihood	Chi-Square	df	Sig.
Intercept Only	13677.353			
Final	5776.207	7901.146	194	.000

Figure 9.4: Model fitting information to represent statistical significance of model

Appendix C

Preprocessing with WEKA

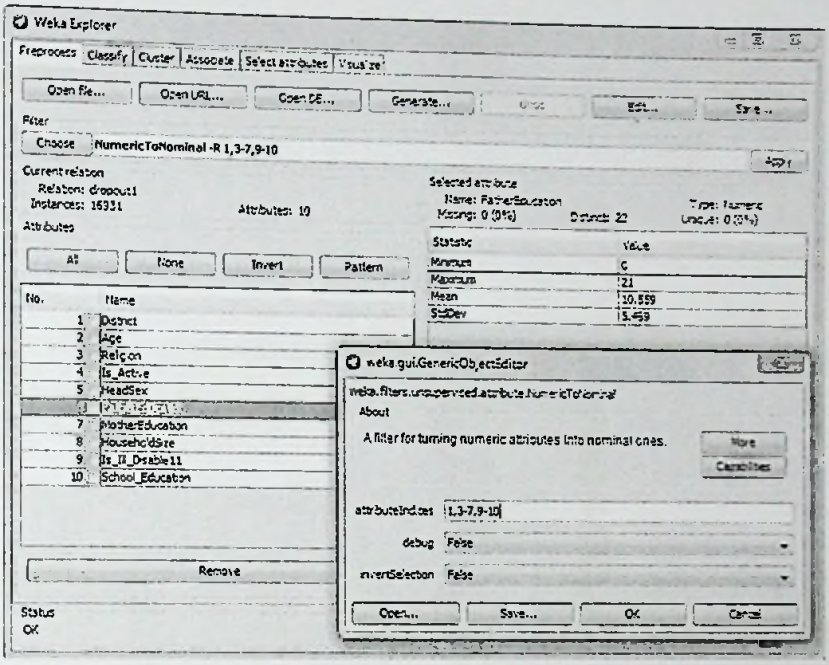

Figure 9.5: Preprocessing with filters

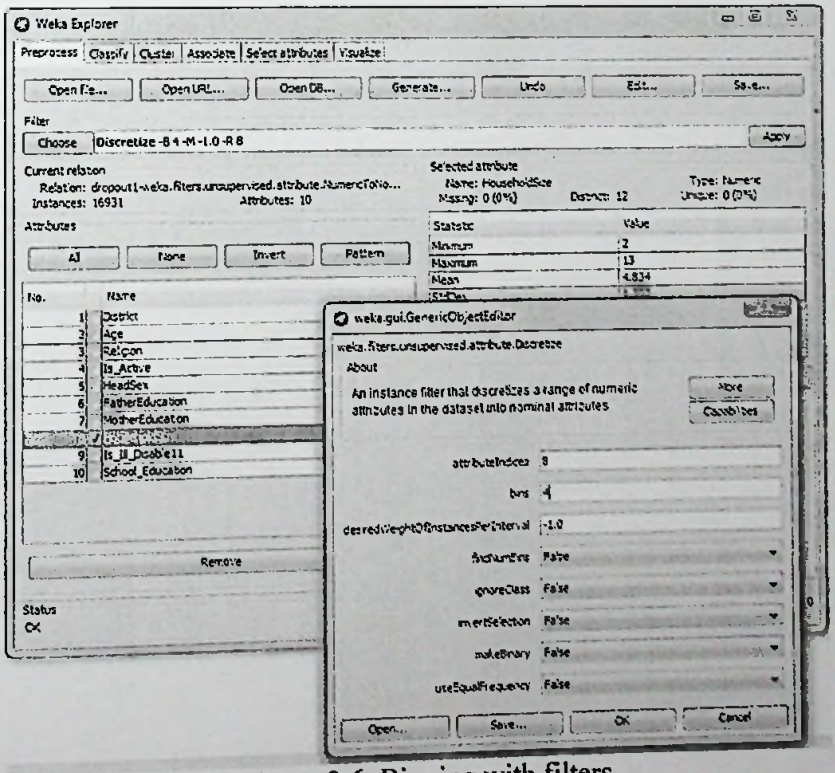

Figure 9.6: Binning with filters

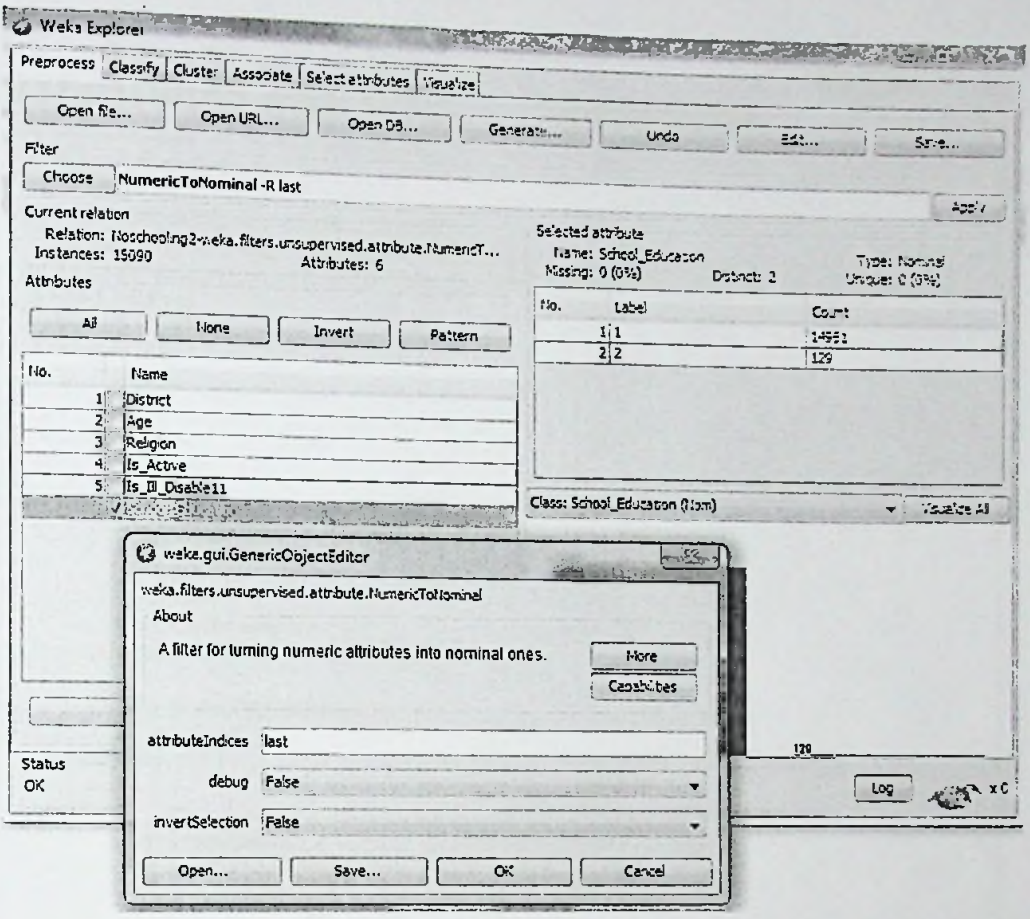

Figure 9.7: Convert class label to nominal for KNN

Appendix D

Data Mining with WEKA-Classification

```
=== Run information ===

Scheme:weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:     dropout1-weka.filters.unsupervised.attribute.NumericToNominal-R1,3-7,9-10-
weka.filters.unsupervised.attribute.Discretize-B3-M-1.0-R2-weka.filters.unsupervised.attribute.Discretize-
B4-M-1.0-R8
Instances:    16931
Attributes:   10
              District
              Age
              Religion
              Is_Active
              HeadSex
              FatherEducation
              MotherEducation
              HouseholdSize
              Is_Ill_Disable11
              School_Education
Test mode:split 66.0% train, remainder test

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

J48 pruned tree
------------------

Is_Active = 1: 3 (709.0/9.0)
Is_Active = 2: 1 (16222.0/1270.0)

Number of Leaves  :         2

Size of the tree :    3


Time taken to build model: 0.31 seconds

=== Evaluation on test split ===
=== Summary ===

Correctly Classified Instances        5350               92.9303 %
Incorrectly Classified Instances       407                7.0697 %
Kappa statistic                          0.5094
Mean absolute error                      0.1374
Root mean squared error                  0.256
Relative absolute error                 67.3125 %
Root relative squared error             81.1504 %
Total Number of Instances             5757

=== Detailed Accuracy By Class ===

                 TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                   1         0.631     0.926     1         0.962     0.684     1
                   0.369     0         1         0.369     0.539     0.684     3
Weighted Avg.      0.929     0.56      0.935     0.929     0.914     0.684

=== Confusion Matrix ===

   a   b   <-- classified as
5112   0 |   a = 1
 407 238 |   b = 3
```

Figure 9.8: Decision Tree for School Dropouts

=== Run information ===

Scheme:weka.classifiers.bayes.NaiveBayes
Relation: dropout1-weka.filters.unsupervised.attribute.NumericToNominal-R1,3-7,9-10-weka.filters.unsupervised.attribute.Discretize-B3-M-1.0-R2-weka.filters.unsupervised.attribute.Discretize-B4-M-1.0-R8
Instances: 16931
Attributes: 10
District
Age
Religion
Is_Active
HeadSex
FatherEducation
MotherEducation
HouseholdSize
Is_Ill_Disable11
School_Education
Test mode:split 66.0% train, remainder test

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

Naive Bayes Classifier

Attribute	Class 1 (0.88)	3 (0.12)
District		
11	1384.0	160.0
12	1233.0	169.0
13	831.0	89.0
21	701.0	92.0
22	404.0	65.0
23	555.0	51.0
31	919.0	133.0
32	845.0	90.0
33	507.0	70.0
41	573.0	82.0
42	261.0	44.0
43	221.0	33.0
44	241.0	42.0
45	393.0	57.0
51	674.0	135.0
52	708.0	96.0
53	478.0	46.0
61	737.0	74.0
62	497.0	71.0
71	530.0	61.0
72	336.0	50.0
81	572.0	65.0
82	396.0	50.0
91	529.0	112.0
92	461.0	58.0
[total]	14986.0	1995.0

Age

'(-inf-10.333333]'	6443.0	17.0
'(10.333333-15.666667]'	5768.0	125.0
'(15.666667-inf)'	2753.0	1831.0
[total]	14964.0	1973.0

Religion

1	8525.0	910.0
2	3087.0	544.0
3	2059.0	312.0
4	1290.0	206.0
9	5.0	3.0
[total]	14966.0	1975.0

Is_Active

1	10.0	701.0
2	14953.0	1271.0
[total]	14963.0	1972.0

HeadSex

1	12579.0	1600.0
2	2384.0	372.0
[total]	14963.0	1972.0

FatherEducation

0	49.0	24.0
1	175.0	51.0
2	319.0	113.0
3	445.0	117.0
4	629.0	135.0
5	1012.0	199.0
6	538.0	118.0
7	744.0	126.0
8	1179.0	151.0
9	793.0	101.0
10	3152.0	248.0
11	1309.0	74.0
12	601.0	23.0
13	1265.0	49.0
14	24.0	1.0
15	261.0	8.0
16	99.0	1.0
17	5.0	1.0
18	3.0	1.0
19	273.0	117.0
20	3.0	1.0
21	2105.0	333.0
[total]	14983.0	1992.0

MotherEducation

0	51.0	15.0
1	121.0	36.0
2	303.0	108.0
3	330.0	86.0
4	530.0	174.0
5	781.0	215.0
6	542.0	137.0

```
7                634.0   124.0
8                843.0   122.0
9                951.0   123.0
10               4172.0  358.0
11               1660.0  101.0
12               772.0   35.0
13               2000.0  55.0
14               32.0    1.0
15               288.0   1.0
16               79.0    1.0
17               3.0     1.0
18               2.0     1.0
19               416.0   186.0
20               4.0     1.0
21               469.0   111.0
[total]          14983.0 1992.0

HouseholdSize
'(-inf-4.75]'    6914.0  802.0
'(4.75-7.5]'     7491.0  1032.0
'(7.5-10.25]'    519.0   133.0
'(10.25-inf)'    41.0    7.0
[total]          14965.0 1974.0

Is_Ill_Disable11
1                457.0   97.0
2                14506.0 1875.0
[total]          14963.0 1972.0


Time taken to build model: 0.04 seconds

=== Evaluation on test split ===
=== Summary ===

Correctly Classified Instances        5320               92.4092 %
Incorrectly Classified Instances      437                7.5908 %
Kappa statistic                       0.5992
Mean absolute error                   0.0949
Root mean squared error               0.2278
Relative absolute error               46.4906 %
Root relative squared error           72.2145 %
Total Number of Instances             5757

=== Detailed Accuracy By Class ===

                 TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.964     0.394     0.951       0.964   0.958       0.948     1
                 0.606     0.036     0.681       0.606   0.642       0.948     3
Weighted Avg.    0.924     0.354     0.921       0.924   0.922       0.948

=== Confusion Matrix ===

   a   b   <-- classified as
4929  183 |   a = 1
 254  391 |   b = 3
```

Figure 9.9: Naïve Bayes Classifier for School Dropouts

```
=== Run information ===

Scheme:weka.classifiers.lazy.IBk -K 1 -W 0 -A "weka.core.neighboursearch.LinearNNSearch -A
\"weka.core.EuclideanDistance -R first-last\""
Relation:     dropout1-weka.filters.unsupervised.attribute.NumericToNominal-Rlast
Instances:    16931
Attributes:   10
              District
              Age
              Religion
              Is_Active
              HeadSex
              FatherEducation
              MotherEducation
              HouseholdSize
              Is_Ill_Disable11
              School_Education
Test mode:split 66.0% train, remainder test

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

IB1 instance-based classifier
using 1 nearest neighbour(s) for classification


Time taken to build model: 0.02 seconds

=== Evaluation on test split ===
=== Summary ===

Correctly Classified Instances        5284               91.7839 %
Incorrectly Classified Instances       473                8.2161 %
Kappa statistic                          0.5823
Mean absolute error                      0.0838
Root mean squared error                  0.2862
Relative absolute error                 41.0367 %
Root relative squared error             90.7335 %
Total Number of Instances             5757

=== Detailed Accuracy By Class ===

                 TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                   0.955    0.38       0.952      0.955    0.954      0.81      1
                   0.62     0.045      0.637      0.62     0.628      0.81      3
Weighted Avg.      0.918    0.342      0.917      0.918    0.917      0.81

=== Confusion Matrix ===

   a   b   <-- classified as
4884 228 |   a = 1
 245 400 |   b = 3
```

Figure 9.10: K-nearest neighbours for School Dropouts

```
=== Run information ===
Scheme:weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:     Noschooling2-weka.filters.unsupervised.attribute.NumericToNominal-R1,3-6-
weka.filters.unsupervised.attribute.Discretize-B3-M-1.0-Rfirst-last
Instances:    15090
Attributes:   6
              District
              Age
              Religion
              Is_Active
              Is_Ill_Disable11
              School_Education
Test mode:split 66.0% train, remainder test

=== Classifier model (full training set) ===
J48 pruned tree
------------------
Is_Active = 1
|   Religion = 1: 1 (8.0/1.0)
|   Religion = 2: 2 (11.0/2.0)
|   Religion = 3: 2 (1.0)
|   Religion = 4: 2 (0.0)
|   Religion = 9: 2 (0.0)
Is_Active = 2: 1 (15070.0/118.0)

Number of Leaves  :         6
Size of the tree :    8

Time taken to build model: 0.14 seconds

=== Evaluation on test split ===
=== Summary ===

Correctly Classified Instances        5079               98.9866 %
Incorrectly Classified Instances        52               1.0134 %
Kappa statistic                          0.101
Mean absolute error                      0.0171
Root mean squared error                  0.0986
Relative absolute error                 96.7818 %
Root relative squared error            100.3851 %
Total Number of Instances             5131

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.999    0.94      0.991     0.999     0.995      0.53     1
                 0.06     0.001     0.375     0.06      0.103      0.53     2
Weighted Avg.    0.99     0.931     0.985     0.99      0.986      0.53

=== Confusion Matrix ===
   a    b   <-- classified as
 5076    5 |    a = 1
   47    3 |    b = 2
```

Figure 9.11: Decision Tree for No schooling

=== Run information ===

Scheme:weka.classifiers.bayes.NaiveBayes
Relation: Noschooling2-weka.filters.unsupervised.attribute.NumericToNominal-R1,3-6-weka.filters.unsupervised.attribute.Discretize-B3-M-1.0-R2
Instances: 15090
Attributes: 6
District
Age
Religion
Is_Active
Is_Ill_Disable11
School_Education
Test mode:split 66.0% train, remainder test

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

Naive Bayes Classifier

Attribute	Class 1 (0.99)	2 (0.01)
District		
11	1384.0	14.0
12	1233.0	10.0
13	831.0	6.0
21	701.0	2.0
22	404.0	5.0
23	555.0	7.0
31	919.0	13.0
32	845.0	13.0
33	507.0	6.0
41	573.0	2.0
42	261.0	1.0
43	221.0	5.0
44	241.0	1.0
45	393.0	2.0
51	674.0	9.0
52	708.0	5.0
53	478.0	5.0
61	737.0	5.0
62	497.0	5.0
71	530.0	6.0
72	336.0	7.0
81	572.0	4.0
82	396.0	9.0
91	529.0	11.0
92	461.0	1.0
[total]	14986.0	154.0

```
Age
 '(-inf-10.333333]'        6443.0   68.0
 '(10.333333-15.666667]'   5768.0   28.0
 '(15.666667-inf)'         2753.0   36.0
 [total]                  14964.0  132.0

Religion
 1                         8525.0   60.0
 2                         3087.0   50.0
 3                         2059.0   14.0
 4                         1290.0    9.0
 9                            5.0    1.0
 [total]                  14966.0  134.0

Is_Active
 1                           10.0   12.0
 2                        14953.0  119.0
 [total]                  14963.0  131.0

Is_Ill_Disable11
 1                          457.0   62.0
 2                        14506.0   69.0
 [total]                  14963.0  131.0

Time taken to build model: 0.06 seconds

=== Evaluation on test split ===
=== Summary ===

Correctly Classified Instances        5081               99.0255 %
Incorrectly Classified Instances        50                0.9745 %
Kappa statistic                          0.1053
Mean absolute error                      0.0165
Root mean squared error                  0.0952
Relative absolute error                 93.5984 %
Root relative squared error             96.8788 %
Total Number of Instances             5131

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.999    0.94      0.991     0.999     0.995      0.829    1
                 0.06     0.001     0.5       0.06      0.107      0.829    2
Weighted Avg.    0.99     0.931     0.986     0.99      0.986      0.829

=== Confusion Matrix ===

   a    b   <-- classified as
5078    3 |   a = 1
  47    3 |   b = 2
```

Figure 9.12: Naïve Bayes Classifier for No Schooling

=== Run information ===

Scheme:weka.classifiers.lazy.IBk -K 1 -W 0 -A
"weka.core.neighboursearch.LinearNNSearch -A \"weka.core.EuclideanDistance -R
first-last\""
Relation: Noschooling2-weka.filters.unsupervised.attribute.NumericToNominal-Rlast
Instances: 15090
Attributes: 6
District
Age
Religion
Is_Active
Is_Ill_Disable1 1
School_Education
Test mode:split 66.0% train, remainder test

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

IB1 instance-based classifier
using 1 nearest neighbour(s) for classification

Time taken to build model: 0.03 seconds

=== Evaluation on test split ===
=== Summary ===

Correctly Classified Instances	5067	98.7527 %
Incorrectly Classified Instances	64	1.2473 %
Kappa statistic	0.1737	
Mean absolute error	0.0155	
Root mean squared error	0.1134	
Relative absolute error	87.9197 %	
Root relative squared error	115.4595 %	
Total Number of Instances	5131	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.996	0.86	0.992	0.996	0.994	0.85	1
	0.14	0.004	0.25	0.14	0.179	0.85	2
Weighted Avg.	0.988	0.852	0.984	0.988	0.986	0.85	

=== Confusion Matrix ===

a	b	<-- classified as
5060	21	a = 1
43	7	b = 2

Figure 9.13: K-nearest neighbours for No schooling

Data Mining with WEKA-Clustering

Appendix E

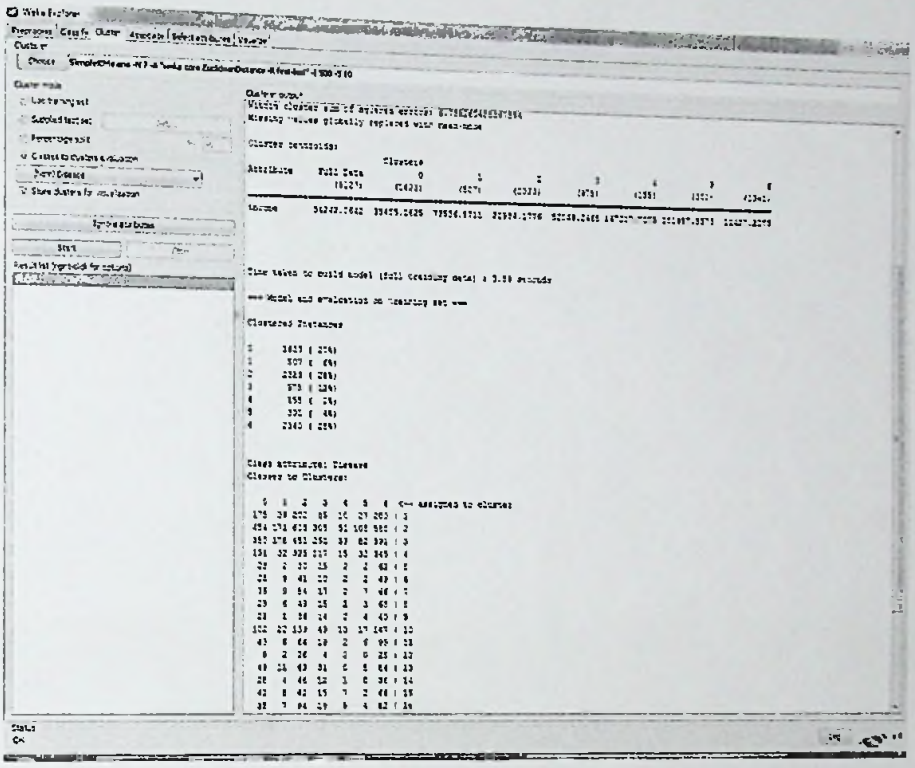

Figure 9.14: KMean clustering Results Window

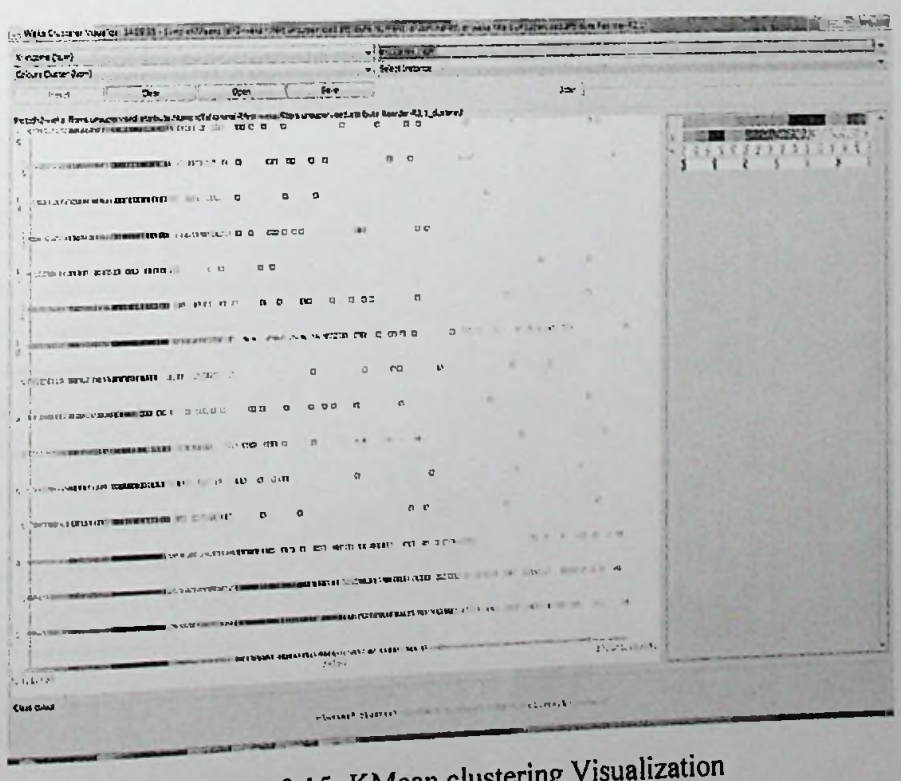

Figure 9.15: KMean clustering Visualization

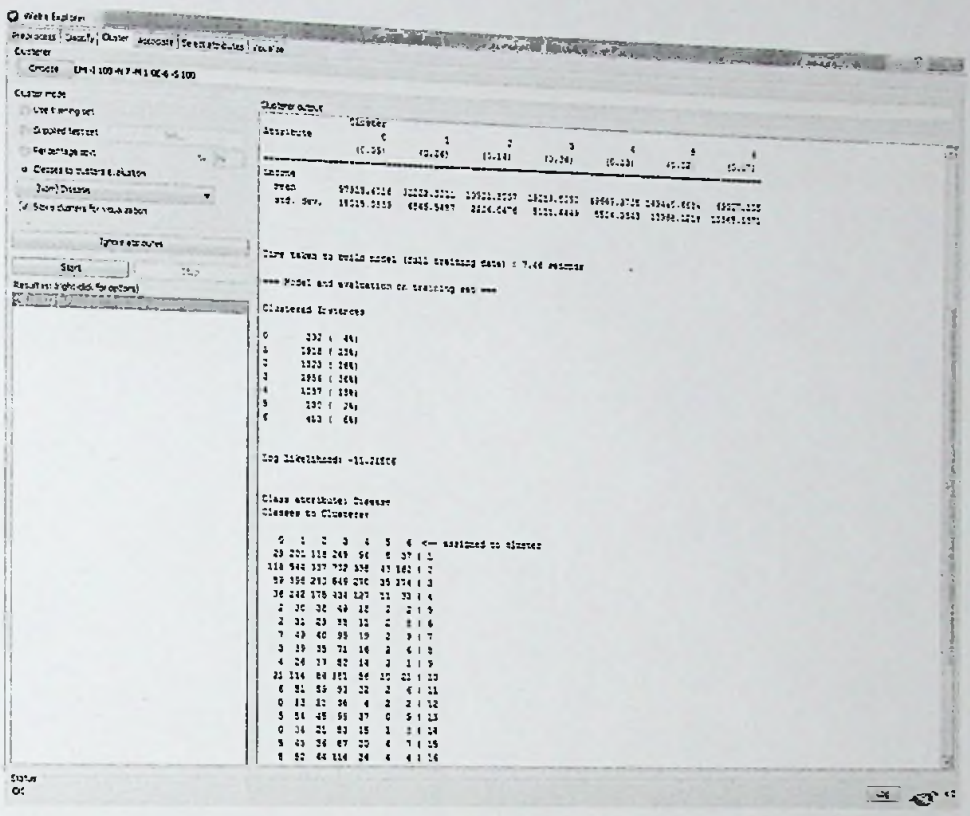

Figure 9.16: EM clustering Results Window

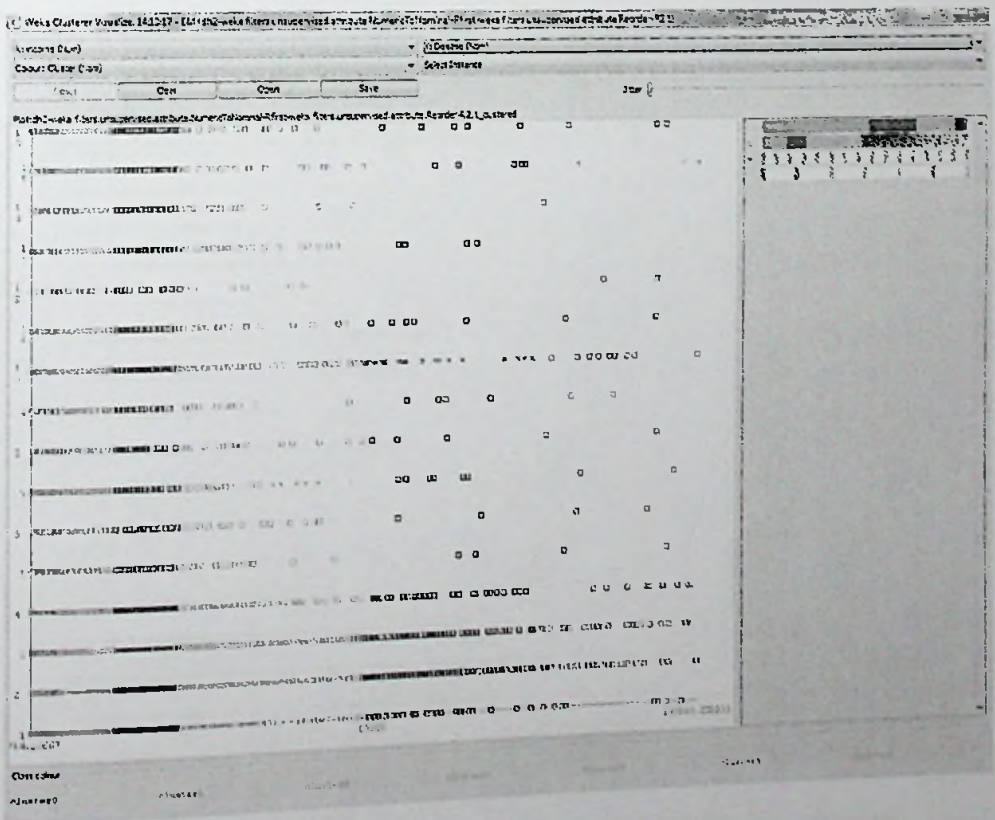

Figure 9.17: EM clustering Visualization

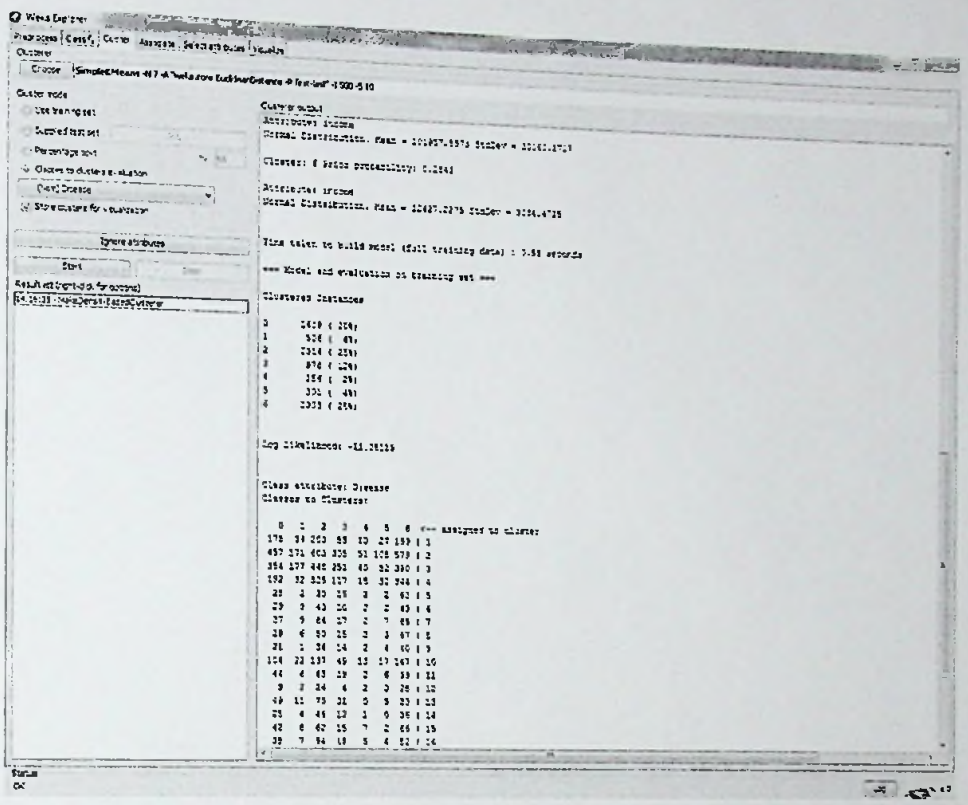

Figure 9.18: MakeDensityBasedClusterer clustering Results Window

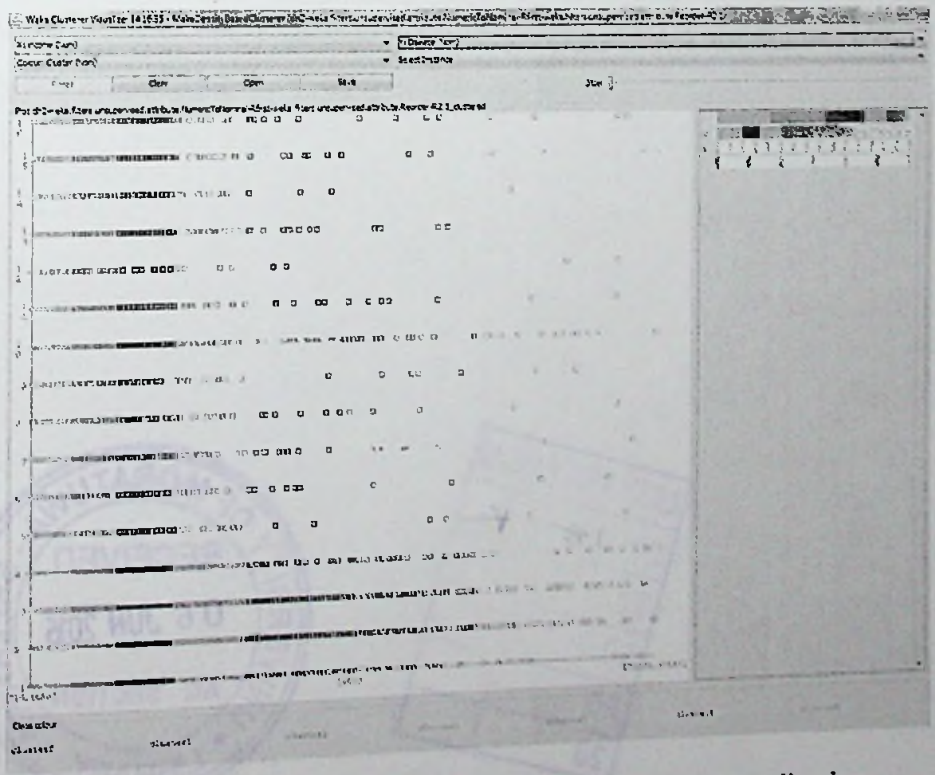

Figure 9.19: MakeDensityBasedClusterer clustering Visualization