Journal of Social and Economic Statistics

Vol. 3, No.2, Winter 2014 *© 2014 Published by JSES.*

DATA MINING EMIGRATION DECISIONS AMONG ROMANIAN TEACHERS. PART 2: THE RESULTS [1]

Angel-Alex HĂISAN[a], Vasile Paul BRESFELEAN[b]

Abstract

The second part of our study analyzes the results obtained from our survey on the teachers' life quality, regarding the factors that influence their decision to emigrate. Based on the methodology described in Part 1, we identified and studied four categories of teachers. For the respondents that have declared their profession income insufficient even for basic necessities, the economic indicators in the decision of emigrating are surpassed by the fulfilment offered by the amount of work they can do and the quality of their work environment. Respondents which declared that profession income covers with great effort their basic necessities display a large array of motivations regarding emigration decision: society and its quality, emotional or financial factors. The teachers from the third category declared their profession income to be just enough for their basic necessities and emigration decision is influenced mainly by their work life. The last category wasn't too relevant for the purpose of this study because only one person from this group wanted to emigrate.

Keywords: teachers, emigration, income, data mining, decision trees
JEL Classification: F22, I31, O15

Authors' Affiliation

[a]Faculty of Economics and Business Administration, Babeș-Bolyai University of Cluj-Napoca, Romania, email: haisan-angel@hotmail.com
[b]Faculty of Economics and Business Administration, Babeș-Bolyai University of Cluj-Napoca, Romania, email: paul.bresfelean@econ.ubbcluj.ro

[1] This paper has been financially supported within the project entitled "Horizon 2020 - Doctoral and Postdoctoral Studies: Promoting the National Interest through Excellence, Competitiveness and Responsibility in the Field of Romanian Fundamental and Applied Scientific Research", contract number POSDRU/159/1.5/S/140106. This project is co-financed by European Social Fund through Sectoral Operational Programme for Human Resources Development 2007-2013. Investing in people!

1. Theoretical and methodological aspects

We start with a brief review of the main theoretical and methodological aspects presented in the first part of this article (Hăisan & Bresfelean, 2014).

1.1. Theoretical aspects

Migration. International migration has important consequences for both sending and receiving countries and its effects represent a hot topic of research (Antman, 2013).

The decline in Romania's population accelerated in the last two decades (Hăisan, 2013), mainly due to emigration to Western European countries and declining birth-rate.

The vast majority of Romanian emigrants is represented by skilled or unskilled workers in domains like agriculture, constructions, textile or extraction industry, but in the last years there has been a high number of highly qualified specialists from medicine, IT or research that left the country seeking a work place abroad (Hăisan, 2013).

Migration of parents abroad for working purposes could influence in a positive way the household income, but can cause irreversible effects on children's well-being and future development (Botezat & Pfeiffer, 2014). The lack of a model for their emotional development frequently causes school dropout, depression or deviant behaviour (Ciuperca, 2009). The ones that are raised by such couples develop disharmonic personalities and may not socially integrate as adult age (Pescaru, 2010).

Researchers have identified, using Maslow's pyramid, the motivations behind emigration wish to be closely correlated to self esteem and they've included here the ones who are looking for a better salary, a higher standard of living or better schools ANBCC (2005).

Data mining. Data mining techniques have known in recent years a widely spread utilization in fields like commerce, marketing, banking, education, medicine, astronomy, etc., because almost every field of humans life has become data-intensive (Venkatadri & Reddy, 2011) and its contribution to decision making processes is invaluable.

The software used was Weka and RapidMiner. Waikato Environment for Knowledge Analysis (Weka) is an open source GNU software developed by the University of Waikato, New Zealand for machine learning (Witten et al., 2011).

Classification learning was the most used data mining method in our research and it allowed us to automatically learn models or rules describing categories of data (Witten et al., 2011). A supervised approach to classification was approached by the use of decision trees, because they could operate under supervision by being provided with the actual outcome for each of the training examples, and the models were used to scan the data and generate the tree and its rules in order to make predictions. These have a "divide-and-conquer" approach and were initially developed for statisticians to automate the process of determining which fields in their database were in fact valuable or correlated with a certain problem (Witten et al., 2011).

Decision trees (Kotsiantis et al., 2006) classify instances by means of sorting them founded on feature values, each node being a feature in an instance to be classified, while each branch is a value that the node can take. Starting with the root node, the instances are classified and then sorted based on their feature values (Kotsiantis, 2007).

The algorithms that generate decision trees have a tendency to automate the hypothesis generation and the validation much more integrated way than any other data mining techniques (Berson et al., 2000). Among their advantages we can include the creation of models that are easy to understand and they are unaffected by missing values in data (Berson, et al., 2000). But they also impose certain restrictions on the analysed, by permitting only single dependent variable (Shah et al., 2006). As a consequence, with the aim of predicting more than one dependent variable, we used separate models for each variable of the distinct groups in our research.

For our classification learning experimentation we've employed J48 and J48graft methods, developed from C4.5 algorithm one of the most used classification algorithms, that offered finer stability between accuracy, speed and results' consistency. J48 represents a greedy algorithm that created decision trees in a top-down recursive "divide-and-conquer" manner. J48graft is an extended version of J48 that considers grafting additional branches onto the tree (Webb, 1999) in a post-processing phase. It tries to attain some of the performance of ensemble methods such as bagged and boosted trees while preserving a sole interpretable structure (Witten et al., 2011).

1.2. Methodology

Our research aims to evaluate the quality of life of Romanian teachers from the North-West region (Hăisan, 2013) and we've built our questionnaire based on the indicators utilized by EQLS (2011), which are referring to: family, economic situation, health, professional life, environment, degree of satisfaction and relationships. The questionnaire was distributed to all physical education teachers from secondary and high schools in Cluj-Napoca, centre of North-West region and a response rate of 70.46% was obtained.

An interesting aspect was observed, after centralizing the data, regarding the "desire to emigrate" indicator (Hăisan, 2013): 38% of the respondents wanted to emigrate and more worrying, 22% of these would emigrate anywhere – which may possibly be interpreted as a desperate gesture.

In consequence, we were set to analyze which could be the indicators that had the most influence in the decision to emigrate and how these decisions differentiate based on the income indicator. In order to achieve this, we've split the study group into four categories by taking into consideration their answers to the income indicator:

1. Their profession income provides them with all the comfort – coded as *"all_confort"* for compatibility with the utilized software;
2. Their profession income covers only their basic necessities – coded as *"basic_necessities"* for compatibility with the utilized software;
3. Their profession income covers with great effort their basic necessities – coded as *"great_effort_basic"* for compatibility with the utilized software;
4. Their profession income doesn't cover even their basic necessities – coded as *"lower_than_basic"* for compatibility with the utilized software.

The fifth category, “NA”, comprised of the ones that haven’t answered to the income question, was excluded because it didn’t bear any relevance.

2. Decision Trees - Interpretation of the results

2.1. First category: “lower_than_basic” - their profession income doesn’t cover even their basic necessities

We start by analysing the first category, namely the ones that have declared that their profession income doesn’t cover even their basic necessities. This group is the smallest from our study, representing 12% of the entire respondents. They are predominantly males, born and raised in an urban area, with the age between 26 and 63 years old. Almost 50% of them are married and have at least one child. More than half of these respondents wish to emigrate.

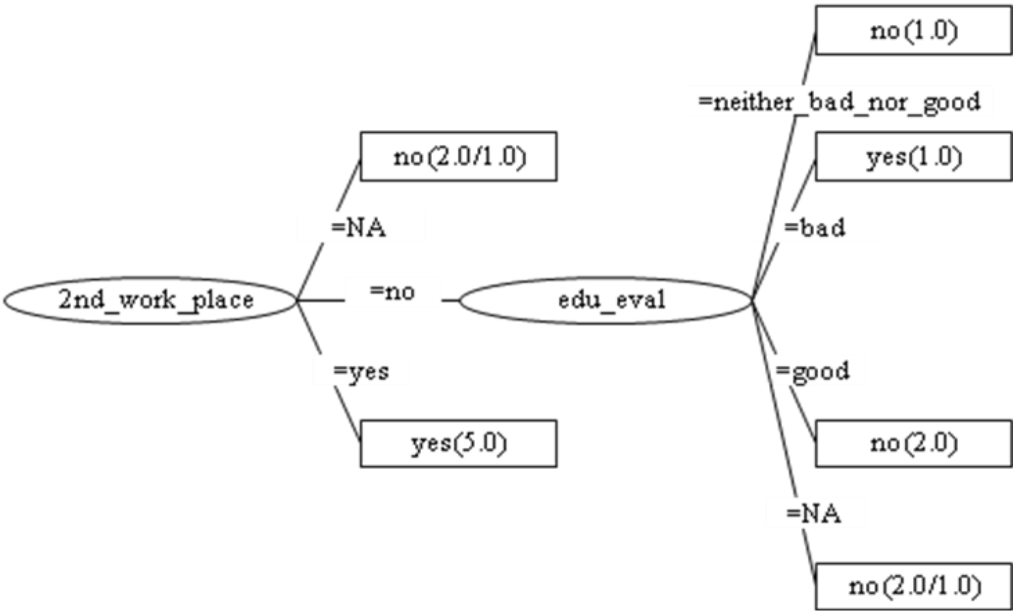

Figure 1 – J48 decision tree based on “lower_than_basic” group

As we can observe from the above figure, for this category the most important indicator in the decision of emigrating or not is the *“2nd_work_place”* indicator. So the ones that haven’t declared anything do not wish to emigrate, while those that have a second job do. This could be explained by the fact that, by having a second job, they are more active and engage more easily in activities, which generates self confidence, so necessarily to adapt and manage in new situations like the emigration process can be.

For those that don’t have a second job, a second indicator is needed in the decision making process of emigration. This indicator refers to the evaluation of the educational system, basically the environment in which they activate. The ones that didn’t respond and those that have evaluated the system as a *“good”* or *“neither_bad_nor_good”* one do not wish to emigrate. The respondents that offered a low score to the educational system would like to emigrate. So we could affirm that the quality of their work environment could be an important factor in the emigration process for our respondents.

The findings in this group correspond to the ones from our other studies in which we've analyzed the will to emigrate based on the marital status of the respondents (Hăisan, 2013).

2.2. *Second category: "great_effort_basic" - their profession income covers with great effort their basic necessities*

We continue by analysing the second largest group in this study, the ones that have declared that their profession income covers with great effort their basic necessities. It represents 33% of the entire study group and is mostly comprised of men, born in an urban area. The age interval of the respondents is between 26-64 years old, with an average of 42. Most of them are married with children and less than half of this group wants to emigrate. The chosen countries are Canada, USA, New Zeeland, Australia or United Arab Emirates.

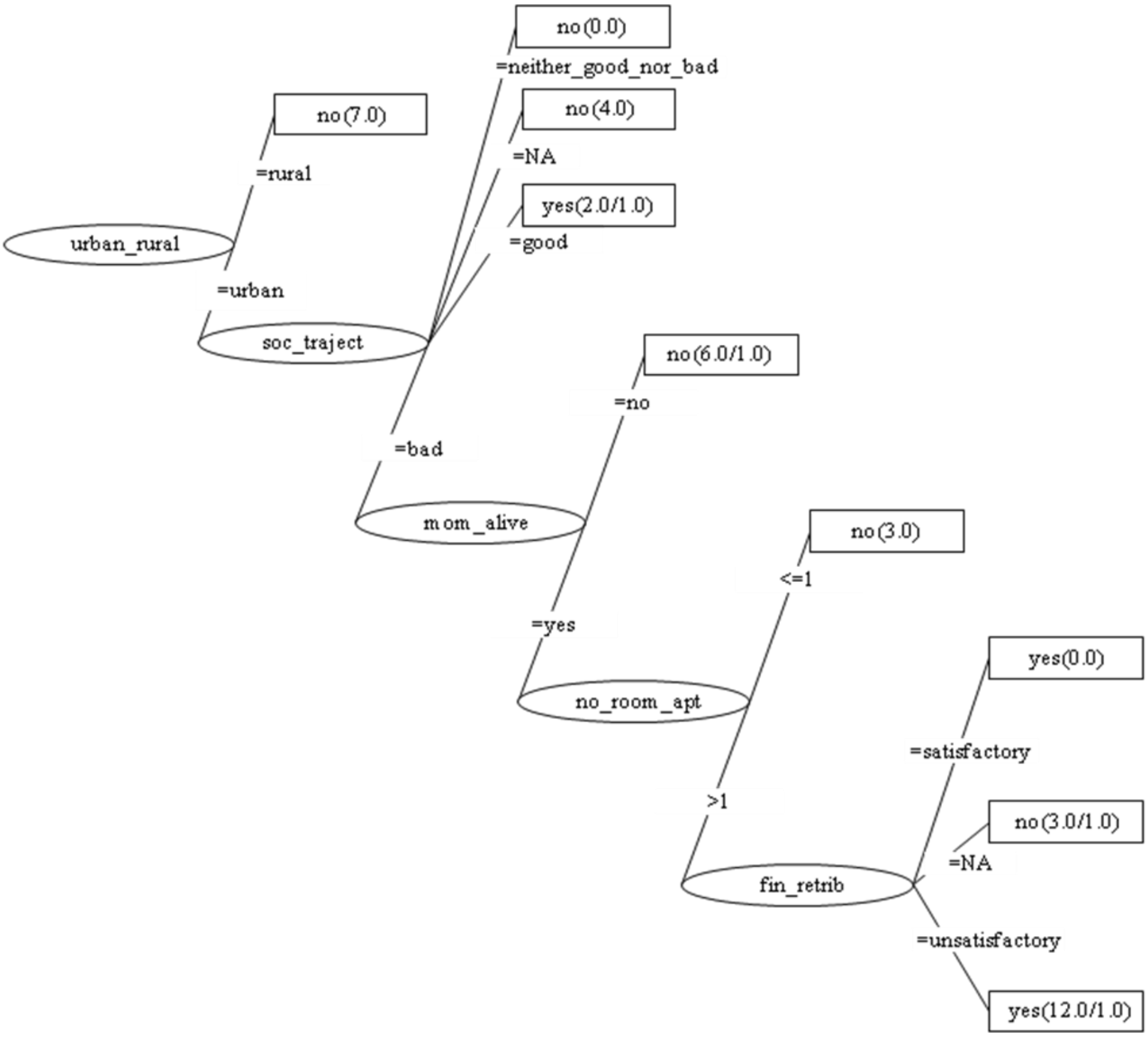

Figure 2 – J48 decision tree based on "great_effort_basic" group

Regarding the indicators that have the biggest influence on the emigration decision of this group's respondents, the first one would be the place where they were born, as we can see from the above figure. If they were born in a rural area they do not want to emigrate, but if they were born in an urban area a second indicator intervenes: the perceived trajectory of the society. So if the respondents were born in an urban area and they consider that the trajectory of the society is *"neither_good_nor_bad"* or didn't express their opinion, they do not want to emigrate. The ones that consider that the trajectory is good surprisingly want to emigrate. The last category, those consider that the trajectory is bad need a third indicator, *"mom_alive"*, meaning if his/hers mom is alive or not. So if a respondent is born in an urban area, considers that the trajectory of the society is bad and his/her mom isn't alive anymore will most likely not emigrate. For those of who their moms are alive a fourth indicator is needed, *"no_room_apt"*, the number of rooms of their apartment. Consequently if a respondent was born in an urban area, considers that the trajectory of the society is bad, his/her mom is still alive and has a one room apartment will not emigrate. If someone is in the same situation as the precedent but has an apartment with more than one room needs a fifth indicator. So lastly if a respondent was born in an urban area, considers that the trajectory of the society is bad, his/her mom is still alive, has an apartment with more than one room and considers that the financial retribution of their job is *"satisfactory"* or *"unsatisfactory"* will most likely emigrate. If he didn't respond to this last indicator it will not emigrate.

2.3. Third category: " basic_necessities" - their profession income covers only their basic necessities

The third group analyzed are the ones that declared their profession income covers only their basic necessities. It is the largest group in this study, representing 35% of the entire respondents and has an almost equal representation between males and females and a very wide age interval 26-69.

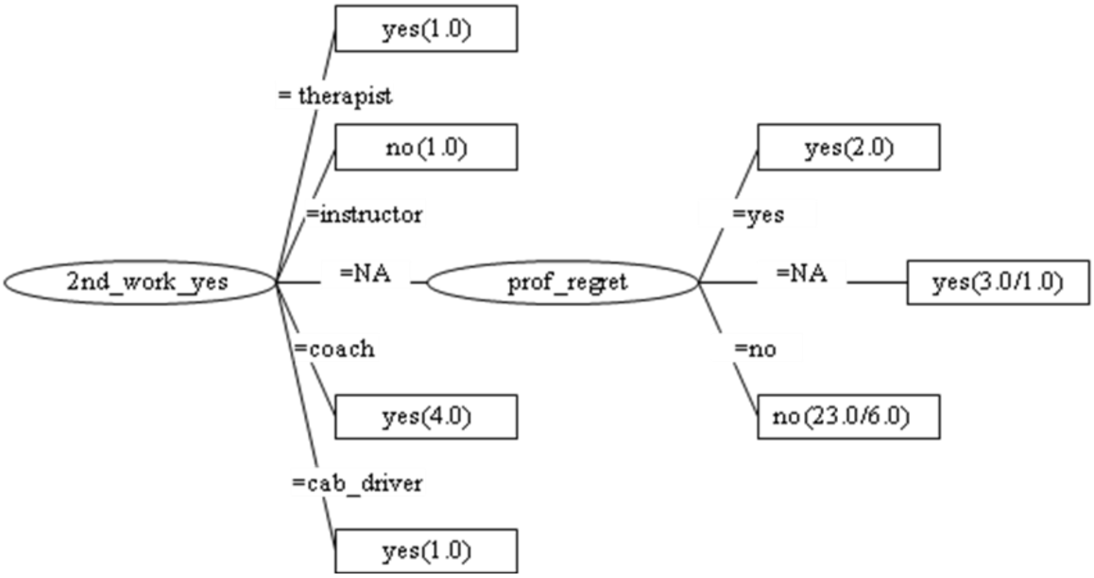

Figure 3 – J48 decision tree based on "basic_necessities" group

The average age is 39 years, more than half being married with children. Most of them are born in an urban area and almost half of them wish to emigrate, preferred countries being USA or "anywhere".

The above figure shows us that the indicator that counts for this category in the emigration process is what type of second job they have. The ones that are cab drivers, coaches and therapists are most likely to emigrate, while instructors won't. A situation somehow similar to the one from the first category analyzed rises in this case also. Most of the *"NA"* cases from the *"2^nd^_work_yes"* indicator mean that the respondents do not have a second job and like in the first case will depend on the work environment. So if they don't regret their profession they will not emigrate, while the ones that regret their choice and those that didn't respond probably will.

2.4. Fourth category: "all_confort" - their profession income provided them with all the comfort

The last category taken into analysis is the one comprised of the respondents that affirmed that their profession income provided them with all the comfort. This group represents the high extreme of our respondents and like the ones that have declared that their profession income doesn't cover even their basic necessities is a small one but not the smallest, which is encouraging. It represents 17% of the entire respondents and is comprised mostly of married men born in an urban area. Most of them have children and their ages are between 26 and 71 years old, with the majority belonging to the 50-70 years interval. Only one respondent from this category wishes to emigrate, which is also the eldest in our study with 71 years.

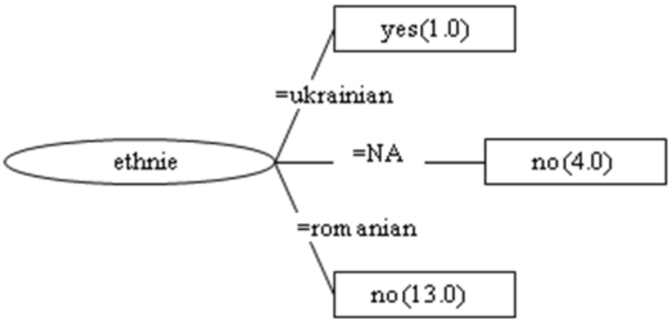

Figure 4 – J48 decision tree based on "all_confort" group

Although not to relevant, because only one persons from this group wants to emigrate, the above figure indicates that for the ones that manage to have everything the indicator that counts is their ethnic group. Interesting enough is the fact that although our respondent is Ukrainian, he doesn't want to immigrate to Ukraine, where presumably has some relatives, but to Canada.

3. Conclusions

The main finding of this ongoing study is identification of the indicators that could influence the decision to emigrate for the four categories of teachers defined in our study.

For the first category, the ones that have declared that their profession income doesn't cover even their basic necessities, an interesting aspect that arises is that although the persons from this category are the ones that struggle financial the most, we would've expect to see some economical indicators in the decision of emigrating, but it seems that these don't weight too much and are surpassed by the fulfilment offered by the amount of work they can do and the quality of their work environment.

The respondents from our second category, *"great_effort_basic"*, display a large array of motivations regarding the decision taking in the emigration process. We have in the first instance motivations related to the society and its quality, then an emotional factor comes into play, the existence or not of the person that gave them birth and finally there are the financial indicators, an indirect one *"no_room_apt"* and a direct one referring to the financial retribution of their job.

For those in the third category, the ones that declared their profession income covers only their basic necessities, the indicators that count in emigrating or not refer to their work life.

The last category, although not to relevant because of the low number of persons that would emigrate, reveals that those that consider that their profession income offers them all the comfort do not take into consideration financial or work related aspects like others do, but more a personal one, that refers to their roots, their ethnic group.

Regarding the method used, we declare ourselves satisfied with the results obtained through the use of it. Data mining aided in establishing raw connection between indicators, based upon which we could continue with a comprehensive interpretation regarding the decision to emigrate.

References

- ANBCC - Asociatia Nationala a Birourilor de Consiliere pentru Cetateni, *Românii si migratia fortei de munca în Uniunea Europeana*, 2005.Bucuresti.
- Antman, F. (2013): The Impact of Migration on Family Left Behind, in: A. Constant, K. F. Zimmermann, International Handbook on the Economics of Migration, Edward Elgar, Northampton, MA.
- Berson A., Smith S. and Thearling K., 2000. *Building data mining applications for CRM*, McGraw Hill, USA.
- Botezat, A., Pfeiffer F., 2014. The Impact of Parents Migration on the Well-Being of Children Left Behind – Initial Evidence from Romania, ZEW Discussion Paper No. 14-029, Mannheim.
- Ciuperca N., 2009. *Efectele emigrarii asupra familiei contemporane*, 1 May, http://ciupercaniculina.blogspot.ro/2009/05/efectele-emigrarii-asupra-familiei.html (accessed August 12, 2014)

- Hăisan, A.-A., 2013. Disfuncționalități în Sistemul Educațional Național – studiu de caz – profesorii din învățământul preuniversitar clujean. Cluj-Napoca: Presa Universitară Clujeană.
- Hăisan A-A, Breșfelean V.P., 2014. Data Mining Emigration Decisions among Romanian Teachers - Part 1: Theoretical and methodological aspects, Journal of Social and Economic Statistics, Vol. 3, No. 1
- Kotsiantis S. B., Zaharakis I.D., and Pintelas P.E., 2006. Machine learning: a review of classification and combining techniques. *Artif. Intell. Rev.*, 159-190.
- Kotsiantis S.B., 2007. Supervised Machine Learning: A Review of Classification Techniques, Informatica 31(2007) 249-268.
- Pescaru M., 2010. *Consecintele migratiei familiei contemporane asupra cresterii si educarii copiilor*, Cluj-Napoca.
- Shah S., Roy R., and Tiwari A., 2006. Technology Selection for Human Behaviour Modelling in Contact Centres, *Decision Engineering Report Series*, Rajkumar Roy and David Baxte (eds.), Cranfield University.
- Venkatadri, M, Loganatha C. Reddy, A Review on Data mining from Past to the Future, International Journal of Computer Applications (0975 – 8887), Volume 15– No.7, February 2011, pp. 19-22.
- Webb G.I., 1999. Decision tree grafting from the all-tests-but-one partition. In Proceedings of the *Sixteenth International Joint Conference on Artificial Intelligence*, 702–707, San Francisco, Morgan Kaufmann.
- Witten I.H., E. Frank and Hall M.A., 2011. *Data mining : practical machine learning tools and techniques.-3rd ed.*, Morgan Kaufmann, Elsevier.
- ****European Quality of Life Surveys* – EQLS, www.eurofound.europa.eu/surveys/eqls/index.htm (accessed August 12, 2014).

Appendix 1 - ***lower_than_basic***

```
=== Run information ===

Scheme:weka.classifiers.trees.J48 -U -M 2 -A
Relation:     lower_than_basic
Instances:    13
Attributes:   27
              income
              marital_stat
              age
              sex
              religion
              ethnie
              urban_rural
              no_child
              mom_alive
              dad_alive
              support_parent
              no_room_apt
              prop_goods
              healt_prob
              memb_home
              marriage
              achivment
              will_emigr
              no_vacations
              2nd_work_yes
              2nd_work_place
              soc_traject
              edu_eval
              prof_regret
              fiz_activity
              fin_retrib
              prof_eval
Test mode:evaluate on training data

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

J48 unpruned tree
------------------

2nd_work_place = no
|   edu_eval = neither_bad_nor_good: no (1.0)
|   edu_eval = bad: yes (1.0)
|   edu_eval = good: no (2.0)
|   edu_eval = N/A: no (2.0/1.0)
2nd_work_place = yes: yes (5.0)
2nd_work_place = N/A: no (2.0/1.0)
```

```
Number of Leaves  :   6

Size of the tree :        8


Time taken to build model: 0 seconds

=== Evaluation on training set ===
=== Summary ===

Correctly Classified Instances          11               84.6154 %
Incorrectly Classified Instances         2               15.3846 %
Kappa statistic                          0.6977
Mean absolute error                      0.2985
Root mean squared error                  0.3339
Relative absolute error                 62.596  %
Root relative squared error             68.5959 %
Total Number of Instances               13

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 1         0.25      0.714     1         0.833      0.95     no
                 0.75      0         1         0.75      0.857      0.95     yes
Weighted Avg.    0.846     0.096     0.89      0.846     0.848      0.95

=== Confusion Matrix ===

 a b   <-- classified as
 5 0 | a = no
 2 6 | b = yes
```

Appendix 2 - *great_effort_basic*

```
=== Run information ===

Scheme:weka.classifiers.trees.J48 -C 0.25 -M 2 -A
Relation:     great_effort_basic
Instances:    37
Attributes:   27
              income
              marital_stat
              age
              sex
              religion
              ethnie
              urban_rural
              no_child
              mom_alive
              dad_alive
              support_parent
              no_room_apt
              prop_goods
              healt_prob
              memb_home
              marriage
              achivment
              will_emigr
              no_vacations
              2nd_work_yes
              2nd_work_place
              soc_traject
              edu_eval
              prof_regret
              fiz_activity
              fin_retrib
              prof_eval
Test mode:evaluate on training data

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

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

urban_rural = urban
|   soc_traject = bad
|   |   mom_alive = yes
|   |   |   no_room_apt <= 1: no (3.0)
|   |   |   no_room_apt > 1
|   |   |   |   fin_retrib = unsatisfactory: yes (12.0/1.0)
|   |   |   |   fin_retrib = N/A: no (3.0/1.0)
|   |   |   |   fin_retrib = satisfactory: yes (0.0)
|   |   mom_alive = no: no (6.0/1.0)
|   soc_traject = good: yes (2.0/1.0)
|   soc_traject = N/A: no (4.0)
|   soc_traject = neither_good_nor_bad: no (0.0)
urban_rural = rural: no (7.0)

Number of Leaves  :    9

Size of the tree :     14


Time taken to build model: 0 seconds

=== Evaluation on training set ===
=== Summary ===

Correctly Classified Instances          33               89.1892 %
```

```
Incorrectly Classified Instances         4               10.8108 %
Kappa statistic                          0.7702
Mean absolute error                      0.2398
Root mean squared error                  0.3001
Relative absolute error                 50.8146 %
Root relative squared error             61.8687 %
Total Number of Instances               37

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.857     0.087      0.857     0.857     0.857      0.949    yes
                 0.913     0.143      0.913     0.913     0.913      0.949    no
Weighted Avg.    0.892     0.122      0.892     0.892     0.892      0.949

=== Confusion Matrix ===

  a  b   <-- classified as
 12  2 |  a = yes
  2 21 |  b = no
```

Appendix 3 - *basic_necessities*

```
=== Run information ===

Scheme:weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:     basic_necessities
Instances:    35
Attributes:   27
              income
              marital_stat
              age
              sex
              religion
              ethnie
              urban_rural
              no_child
              mom_alive
              dad_alive
              support_parent
              no_room_apt
              prop_goods
              healt_prob
              memb_home
              marriage
              achivment
              will_emigr
              no_vacations
              2nd_work_yes
              2nd_work_place
              soc_traject
              edu_eval
              prof_regret
              fiz_activity
              fin_retrib
              prof_eval
Test mode:evaluate on training data

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

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

2nd_work_yes = N/A
|   prof_regret = no: no (23.0/6.0)
|   prof_regret = N/A: yes (3.0/1.0)
|   prof_regret = yes: yes (2.0)
2nd_work_yes = coach: yes (4.0)
2nd_work_yes = cab_driver: yes (1.0)
2nd_work_yes = instructor: no (1.0)
2nd_work_yes = therapist: yes (1.0)

Number of Leaves  :   7

Size of the tree :        9


Time taken to build model: 0 seconds

=== Evaluation on training set ===
=== Summary ===

Correctly Classified Instances          28               80      %
Incorrectly Classified Instances         7               20      %
Kappa statistic                          0.5868
Mean absolute error                      0.2915
Root mean squared error                  0.3818
Relative absolute error                 58.7103 %
Root relative squared error             76.6371 %
Total Number of Instances               35

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.947     0.375      0.75      0.947     0.837      0.809    no
                 0.625     0.053      0.909     0.625     0.741      0.809    yes
Weighted Avg.    0.8       0.228      0.823     0.8       0.793      0.809

=== Confusion Matrix ===
  a  b   <-- classified as
 18  1 |  a = no
  6 10 |  b = yes
```

Appendix 4 - *all_confort*

```
=== Run information ===

Scheme:weka.classifiers.trees.J48 -U -M 2
Relation:     all_confort
Instances:    18
Attributes:   27
              income
              marital_stat
              age
              sex
              religion
              ethnie
              urban_rural
              no_child
              mom_alive
              dad_alive
              support_parent
              no_room_apt
              prop_goods
              healt_prob
              memb_home
              marriage
              achivment
              will_emigr
              no_vacations
              2nd_work_yes
              2nd_work_place
              soc_traject
              edu_eval
              prof_regret
              fiz_activity
              fin_retrib
              prof_eval
Test mode:evaluate on training data

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

J48 unpruned tree
------------------

ethnie= romanian: no (13.0)
ethnie= N/A: no (4.0)
```

```
ethnie= ukrainian: yes (1.0)

Number of Leaves  :   3

Size of the tree :        4


Time taken to build model: 0 seconds

=== Evaluation on training set ===
=== Summary ===

Correctly Classified Instances          18              100      %
Incorrectly Classified Instances         0                0      %
Kappa statistic                          1
Mean absolute error                          0
Root mean squared error                          0
Relative absolute error                      0      %
Root relative squared error                    0      %
Total Number of Instances                 18

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 1         0          1          1        1          1        nu
                 1         0          1          1        1          1        da
Weighted Avg.    1         0          1          1        1          1

=== Confusion Matrix ===

  a  b   <-- classified as
 17  0 |  a = no
  0  1 |  b = yes
```