Journal of Social and Economic Statistics

Vol. 3, No.1, Summer 2014 *© 2014 Published by JSES.*

DATA MINING EMIGRATION DECISIONS AMONG ROMANIAN TEACHERS - PART 1: THEORETICAL AND METHODOLOGICAL ASPECTS[1]

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

Abstract

Family is considered to be the foundation of any society. In former communist countries from Eastern Europe there have been serious concerns that slowly deteriorate this status: emigration issues, increased divorce rate, declining marriages rate, increased mortality and low life expectation etc. All these outline some crucial threats for the survival of a nation. Since the research literature offers few surveys on the teachers' life quality, we've conducted a study on teachers from Cluj-Napoca, the second largest city in Romania. We've collected various indicators, referring to life standards, job, income, emigration, opinions and discontent, general and particular aspects of life and career, in order to find out of sight connections using data mining techniques. The decision trees algorithms were used to establish raw connections between indicators and to formulate a comprehensive interpretation on the emigration's influence factors. The results of this study will be presented in an upcoming article (Part 2).

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

Authors' Affiliation

[a]Faculty of Economics and Business Administration, Babeş-Bolyai University of Cluj-Napoca, Romania, email: haisan-angel@hotmail.com
[b]Faculty of Economics and Business Administration, Babeş-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!

"A great emigration necessarily implies unhappiness of some kind or other in the country that is deserted."

Thomas Malthus

1. Introduction

It is generally acknowledged (Antman, 2013) that international migration has important consequences for sending countries (typically part of the developing world) as well as for receiving countries, and the direction and magnitude of these effects are increasingly investigated. Romania has known in the last decade a tremendous diminution of its population (Hăisan, 2013), mainly due to immigration to Western European countries and declining birth-rate. According to the national census (CCRPL, 2012), approximately 910.264 persons legally left the country for more than 12 months, with the purpose of finding a job, for studies or business. Unfortunately, experts' estimations seems to be more pessimistic (Mihai, 2011), mentioning a number of 2.1 million Romanians who left the country, many of them belonging in the 20-35 years old category, while over 60% were women (Pritulescu, 2011).

Parents' migration abroad for working purposes may have some positive effects as a way of generating income, but on the other hand, missing the main adult caregiver may be harmful for children's well-being and future development (Botezat & Pfeiffer, 2014). Family represents "the fundamental cell of society where we learn to live with others despite our differences and to belong to one other" (Pope Francis), and the wellbeing of each cell can lead to the wellbeing of the whole body. Sadly, Romanian emigration has affected in a major way the families and the communities in which they live, thus shaping family life in our country.

New studies (Antman, 2013), (Botezat & Pfeiffer, 2014) concentrate on the migration's effects on the separation of families, whether it is a small part of the family separating from the extended family or a parent migrating alone. In view of that, the consequences of emigration on families are most severe, and the worst case scenario is when a mother leaves her family and children (Hăisan, 2013), thus leading in many cases to a family scission and the alienation of its members. This type of migration is considered to be circular and recurrent, also determining specific questions about the impact of migration on family members left behind, their dependence on the emigrant for sustenance but not only that (Antman, 2013), (Botezat & Pfeiffer, 2014). Unfortunately, children coming from with this kind of background may grow up deprived of indispensable human touch and emotions to their evolution as individual, while children raised by both parents beside them would face a more favourable and wellbeing state (Hăisan, 2013).

The bulk of Romanian emigrants are represented by skilled or unskilled workers in certain domains (agriculture, constructions, textile or extraction industry) but, especially in the last years, there have been high qualified specialists in fields such as medicine, IT, research who left the country to work abroad (Hăisan, 2013). An important percentage of young alumni from Romanian medical schools and faculties expressed their will for a better life outside the country (Moldoveanu, 2011).This is a major loss in a country with a dominant state owned and mostly free education, where the government spends an average of approximately 10.000 Euros for each citizen until employment and doubles the sum for college graduates (Cuncea, 2012). Since continuing changing governments have not been able to prove a finality of the educational process nor a vigorous society with real facilities and competitive salaries, the state deflects its alumni it invested in towards other countries. This situation is also reflected in the poor results obtained at the national baccalaureate, with a gloomy graduation rate of less than 50% in the last years. Some of the factual roots of this can be distinguished as follows (Hăisan, 2013): less and less prepared students and teachers, numerous changes in the system, while teachers who have dedicated their lives to this profession struggle each day for subsistence and dignity.

Romanian central administration decided in 2007 to exclude physical education and sport from the national baccalaureate examinations, which had a grave consequence: it lowered the importance and the number of hours allocated to these classes, and placed them into an "etcetera" category. This was an unfortunate measure, since it is generally agreed (Bailey, 2006), (Warburton et al., 2006) that sport and sport teachers have a vital mission for children, to induce the fondness for exercise and lay the foundation for a good state of health. Furthermore, sport could help develop a healthy body and mind, represent a prophylactic treatment for various diseases, offering the lucidity needed to face daily problems (Hăisan, 2013), like the old Latin axiom once stated: "Mens sana in corpore sano".

An important part of our research has been committed to investigating the factors that lead to Romanian teachers' wish to emigrate, but we also sought to identify the factors that lead them to consider the educational system to be of low quality (Hăisan, 2013). Furthermore, we targeted an important part of them, namely the physical education and sport teachers from secondary and high schools and used the data mining specific decision trees algorithms to establish raw connections between indicators and to formulate a comprehensive interpretation on the emigration's influence factors.

2. Problem Formulation

In our research made to evaluate the quality of life of Romanian teachers from the North-West region (Hăisan, 2013), we've included several survey indicators also utilized by EQLS (2011) referring to: family, economic situation, health, professional life, environment, degree of satisfaction and relationships. We've distributed questionnaires to all physical education teachers from secondary and high schools in Cluj-Napoca, centre of North-West region. Conclusively, we obtained a 70.46% response rate, since some teachers declined the invitation to participate to our study.

An interesting result that was obtained after tabulating the responses, regarded "desire to emigrate" indicator (Hăisan, 2013), with a prominent percentage in favour of emigration (38%), while 22% of these particular respondents declared they would emigrate "anywhere" – which may possibly be interpreted as a desperate gesture. Unfortunately, these high numbers are not a novelty among qualified specialists within Romanian society and should be an alarm for the national and European authorities.

As a result, 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.

In our main study, from which we've extracted the emigration indicator (Hăisan, 2013), we have managed to obtained data referring to more aspects of teachers' life, not only to personal and financial aspects and therefore we could generate specific connections. In order to achieve that, we've employed data mining techniques aiming to establish a connection between indicators that at a first look do not seem to influence each other.

In Data Mining an important assignment (Quinlan, 1993) is to generate models in order to predict the class of an object on the basis of its attributes. In our case the main study objects (subjects) are represented by teachers, with attributes related to their age, gender, location, religion, marital status, family, income, health, job, environment, several opinions, satisfaction and discontent for different aspects of life and career (etc.), whereas the class of the subjects would be positive or negative for their desire to emigrate. In the present paper we consider the problem of learning classification tree models using the collected data from the teachers in Cluj-Napoca, the second largest city in Romania.

3. State of the Art Research

3.1. Emigration, Family and Education

The effects of parental migration on children were debated in various studies. It was generally agreed that some of the most serious effects emerged when the mother left home. The effects are felt both by partners but especially by children, left without parental affectivity. Other effects on children could be represented by the lack of a model for their emotional development that frequently causes school dropout, depression (leading to suicide) or a deviant behaviour (Ciuperca, 2009). Children raised by such couples could develop disharmonic personalities and may not socially integrate as adults (Pescaru, 2010).

In the study entitled "The impact of the Economical Crisis on the Migration of the Romanian Workforce", the authors showed that the will to emigrate did not depend any longer on age, ethnicity, profession or on the separation from the close ones that had already emigrated (Stanculescu, 2011). For instance, the oldest respondent in our study that wanted to emigrate was 71 years old. The same study (Stanculescu, 2011) showed that in 2010, the emigrants had completed the following levels of education: primary, secondary and vocational (47.4% of them); high school and post high school (9.2%); higher education (13.4%).

Regarding the principal motivations behind emigration wishes, ANBCC (2005) conducted a study entitled "Romanians and the migration of workforce in European Union" using Maslow's pyramid. A large majority of the subjects (75%) were placed on the 4th step of the pyramid, namely self esteem. Researchers from ANBCC included in this category individuals searching for a better salary (51.49%), for a higher standard of living (20%), for better schools, for short work related visits and for learning a new language (these last three totalizing 2,32%). The self esteem issues were followed by the need for safety, the need for accomplishment and the need of social affiliation (ANBCC, 2005).

Other studies were conducted on the quality of the Romanian educational system, and regardless the positions of the ones involved in the system (teachers or students), the conclusions were the same: the quality of the Romanian educational system was poor and it had been experiencing a constant degradation (Nedelcu, 2010). In 2010 ICCV studies publicized similar conclusions on the educational system. The educational system was perceived by 32% of the respondents to be of low and very low quality, an increased percentage compared to the results obtained in 2006 by the same study regarding the quality of life of Romanian citizens (Marginean et al., 2010). In terms of quality of life, the same study showed that 74% of the population felt the living conditions were worst than the year before (Marginean et al., 2010).

3.2. Data Mining Experiments in the Field

In recent years, data mining technologies have known broad application, from commerce, marketing, banking, to education, medicine, astronomy, etc., due to its importance in decision making processes and programs within various institutions. Owing to the fact that almost every field of human life has become data-intensive (Venkatadri & Reddy, 2011), data mining technologies were extended into new areas of human life with numerous integrations and advancements in fields like: Statistics, Databases, Machine Learning, Pattern Reorganization, Artificial Intelligence and Computation capabilities etc.

These technologies have emerged due to the continuing increase of databases' development and sophistication. From the research literature (Ionita, 2005), we present a brief history of these developments in modern databases:

- 1960s - the collecting of data in various formats digitization which allowed a retrospective analysis of data by computer;
- 1980s - relational databases appeared along with the (Structured Query Language (SQL), thus allowing dynamic analysis of data by request.
- 1990s - were characterized by an explosion of data, and by the appearance of data warehouses.

According to academy professor Florin Filip (2000), data mining, also known as knowledge discovery in databases (KDD) is considered to be a data analysis technology, associated with OLAP, and with the data warehouse concepts. Some of the definitions given to these technologies from the research literature (Filip, 2000) are as follows:

- An older definition was given by (Frawley, 1991, cited by Mertens et al, 1996) which stated that data mining was a not so simple and trivial extraction of potentially useful information, implied and acknowledged before a database.
- Fayyad et al. (1996) defined KDD as the "nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data", turning disordered data into valuable knowledge. The steps for KDD included: data cleaning, data reduction, data transformation, data mining, pattern evaluation and then knowledge discovered.
- Other researchers (Moxon, 1996) believed that data mining was a set of techniques used to address automated operations of exhaustive exploration in large databases and bring up the complex relationships existing in those databases. Professor Ullman from Stanford University InfoLab (cited by Breşfelean, 2008), affirmed that data mining primarily meant the overuse of statistical data to deduce invalid inferences.

One of the most used methodology in the last decade for data mining and knowledge discovery has been CRISP-DM (CRoss Industry Standard Process for Data Mining) developed by a group of prominent enterprises (Teradata, SPSS-ISL, Daimler-Chrysler, OHRA). It acted as vendor-independent with the possibility of being used with any data mining tool and solve any data mining problem (Chapman et al., 2000), (Marban et al., 2009), by defining the phases to be carried out, the tasks and the deliverables for each task. CRISP-DM (www.crisp-dm.org) (Chapman et al., 2000) as cited by (Bresfelean, 2009) was divided into six phases:

1. Business Understanding (in our case the research understanding) - centers on understanding the project objectives and necessities from a business perspective, and then transforms the knowledge into a data mining problem definition.
2. Data Understanding - starts with an initial data collection and proceeds with activities in order to get familiar with the data, to identify data quality problems, to discover first insights into the data, or to detect interesting subsets.
3. Data Preparation - consists of all activities to construct the final dataset from the initial raw data.

4. Modeling - various modeling techniques are selected and applied, and their parameters are calibrated to optimal values.
5. Evaluation - the user has built a model that appears to have high quality, from a data analysis perspective, wand at the end of this phase, a decision should be attained regarding the use of the data mining results.
6. Deployment - Creation of the model is generally not the end of the project. Depending on the requirements, the deployment phase can be as simple as generating a report or as complex as implementing a repeatable data mining process, expanding the obtained model and its results at the level of managerial information system of the higher education institution.

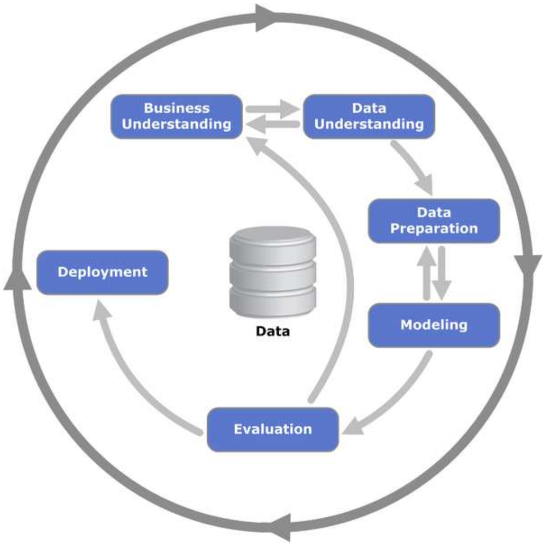

Figure 1 - The 6 phases of the CRISP-DM model (Chapman et al., 2000)

In our documentation phase for our experiment, we’ve tried to find similar studies in the research literature, involving studies on teachers, family, income, but also a number of implications of social life and living standards, such as emigration. We found a series of very interesting articles, but they did not deal with this large family trend, its causes and implications. Literature studies treat only parts of the phenomenon in innovative ways, some of them being presented in the following paragraphs. It is important to remark that most of the research is based on existing data and information from various databases and the Internet, and does not try to collect data directly from the source, for example from teachers and their families, in an attempt to better perceive their opinions, lifestyle, dissatisfactions and wishes. This might be more difficult to accomplish, due to the fact that it involves interdisciplinary research, a fusion of sociology, computer science, statistics etc., each backed by an expert in the field. It is probably handier to researchers to use data from national or international statistic institutes, e-learning systems, management systems, social networks etc.

In the research literature we could find few applications of data mining in traditional education, but many in e-learning and web-based adaptive education systems. An "educational data mining" term is mentioned in some papers, as a set of data mining techniques with applications in most educational phenomena, with emphasis on data obtained from log files and databases from E-learning and Learning Management Systems, communication, collaboration, administration and reporting tools etc. The resulted information and knowledge from these processes can be used by teachers and courses' authors to improve their performance, as well as by the education institutions' managers for decision making processes (Bresfelean, 2008), to learn from students content/discontent regarding curricula, courses, exam failure etc.

Romero et al., (2005) used evolutionary algorithms as data mining methods to improve courseware, especially Adaptive Systems for Web-based Education (ASWE). They collected data during students' sessions which could be useful for the teachers and courses' authors, and used Grammar-Based Genetic Programming with multi-objective optimization techniques. Then, they could specify the form of the rules and choose several evaluation measures of the rules' quality with the purpose to develop a specific mining tool for discovering rules and improving ASWE methodologies.

Garcia et al., (2009) continued the research and proposed a system oriented to discover, share and suggest suitable changes to improve courses' effectiveness. In one of their articles they described an iterative methodology to develop and carry out the maintenance of web-based courses to which we have added a specific data mining step. They also used a collaborative recommender system to share and score recommendation rules obtained by teachers and experts in education, and conducted experiments with several groups of students.

Other important applications of data mining are related to the social networks such as Facebook, Twitter, LinkedIn, etc., which generate a remarkable quantity of data. Russell's (2011) book shows several ways to combine social web data, analysis techniques, and visualization, using techniques for mining data in different areas of the social Web, including blogs and email. Some of the methods proposed in the book, include: basic Python tools, adaptable scripts on GitHub, XHTML, TF-IDF, cosine similarity, collocation analysis, document summarization, clique detection, interactive visualizations with web technologies based upon HTML5 and JavaScript toolkits etc. Eagle and Pentland (2005) tried to identify social networks' evolution, the predictability in people's lives and the information flows, naming the method "reality mining".

There have been developed certain "social data mining systems" which could enable people to share opinions and benefit from each other's experience. Amento et al. (2003) presented TopicShop, a system that mined information from the structure and content of Web pages and offered an exploratory information workspace interface. It used and redistributed information from computational records of social activities, messages, system usage history, citations, or hyperlinks. TopicShop interface allowed users to select considerably more high-quality sites, in less time, and to organize them into personally significant collections. Mikheyenkova (2011) wrote about an "intellectual data mining", a combination of qualitative and qualitative methods of social data mining in combination with subjective research methods. The author tried to prove the need for the intellectualization of such analysis using methods of modern computer intellectual systems, concluding that even though they cannot substitute the work of a sociologist, they extend human abilities to process empirical social data and contribute to the objectification of social studies.

Annoni et al. (2006) conducted data mining studies on Italian families' behaviour and expenditures on durable and daily goods and on various services, using yearly data from Italian National Bureau of Statistics (ISTAT). Their purpose was twofold: describing the most important characteristics of family behaviour with respect to expenditures on goods and usage of different services; highlight and explain possible relationships among these behaviours by social-demographical features of the families.

While the Internet abounds with opportunities for learning, communicating, and sharing information, it also has various risks, especially for children. As a consequence, parents must be alert to potential problems their children might encounter, and take measures to protect them. Lin and Shih (2008) explore some interesting correlations between: parents' information literary, children's ability of self-defence on the Web, and adequate measures to encourage them to use the Web more effectively. They applied association rules, based on Apriori classic algorithm for learning association rules, which was designed to function on databases containing transactions.

Founded on 7 years of examining data from 5,350 male employees (40-49 years old group), Ogasawara et al. (2006) applied association rule analysis to elucidate the relationship between 6 lifestyles (overweight, drinking, smoking, meals, physical exercise, sleeping time, and meals), 5 family medical histories (hypertension, diabetes, cardiovascular disease, cerebral-vascular disease, and liver disease), and 6 medical abnormalities (high blood pressure, hyperchoresterolemia, hypertrigriceridemia, high blood sugar, hyperuricemia, and liver dysfunction). They established that the data mining resulted groupings were superior to the ones derived from traditional methods (logistic regression analysis), and that association rule methods proved more valuable to clarify combinations of risk features for lifestyle diseases.

Data mining also has applications in the family tree: identification of relatives using genetic kinship analysis of DNA. Beiber et al. (2005) examined the possible use of DNA databases to look for perpetrators who are related to those in the lawbreaker DNA database. Forensic kinship analysis was developed based on traditional paternity testing, in which probability calculations were made to determine genetic relationships between individuals based on DNA, and were also used for identification of mass disaster victims. To enhance the success of their methods, the authors propose search combinations with other governmental data: geographic information systems data, demographic and genealogical data.

The City University of Hong Kong developed an e-Government project (Chun, 2007) with artificial intelligence services to support automated assessment of data from immigration agency, developed by. The services were integrated into the agency's application form processing system and included: data mining and machine learning, rule-based assessment, workflow processing, schema-based suggestions, case-based reasoning. The objective was to provide faster and higher quality service to citizens and visitors in fairly and accurately processing their requests while considering all relevant laws and regulations.

3.3. Classification Learning Methods

Data mining has been a pioneering field of research that started as an effect of explosive increase in the amount of stored data from all aspects of socio-economic life. It includes a set of wide-ranging analytical tools for the operation of very large databases and the discovery of complex associations and knowledge, without previous hypotheses formulation. Data mining answers to diverse research questions and has a constructive utility in decision-making processes.

In our research we've employed a couple of data mining software, such as Weka and RapidMiner. Weka 3 (Waikato Environment for Knowledge Analysis) is a popular open source GNU software for machine learning that has been developed at the University of Waikato, New Zealand (Witten et al., 2011). It includes a large collection of machine learning algorithms for data mining tasks, which can be directly utilized to a dataset or call from users' Java code. Weka has tools that can be used for various processes, such as: data pre-processing, classification, regression, clustering, association rules, visualization and to develop new machine learning schemes.

RapidMiner is another world-leading open-source system for data mining, available as a stand-alone application for data analysis and as a data mining engine for the integration into own products. Some of RapidMiner's most important features include: data integration, analytical ETL, data analysis, repositories for process, data and meta data handling, data transformation, data modelling, and data visualization methods, on-the-fly error recognition and quick fixes etc.

Basically, there are considered to be four main categories of learning in data mining applications (Witten et al., 2011), as follows:

1. Classification learning - where the learning scheme is presented with a set of classified examples from which it is expected to learn a way of classifying unseen examples.
2. Association learning - which seeks for all associations among features.
3. Data clustering – where groups of examples that belong together are sought.
4. Numeric prediction – where the outcome to be predicted represents a numeric quantity.

One of the prominent data mining methods that was used in our research, is the classification learning that allowed us to automatically learn models or rules describing categories of data (Witten et al., 2011). We've relied on a supervised approach to classification, namely 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. Decision trees 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, and represent "divide-and-conquer" approach (Witten et al., 2011) to the problem of learning from a set of instances.

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 then 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.

Some of the most used decision trees algorithms include: CART (Classification And Regression Trees), CHAID (Chi-squared Automatic Interaction Detection), Quest, ID3 (Iterative Dichotomiser 3), C4.5, C5.0 etc. Decision trees are generated using data iterative splitting into distinct groups, with the purpose to maximize the groups' distance at every split. Leaf nodes offer a classification applied to all instances reaching the leaf (Witten & Frank, 2011) or a set of classifications, or a probability distribution over all possible classifications.

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 structur (Witten et al., 2011).

In the present paper, in an attempt to forecast the teachers' wish to emigrate, we present the resulted decision trees based on the J48 method, where we've applied a pruning technique that replaced subtrees with leaves in an attempt to reduce over-fitting and also the Laplace estimator that commenced all numbering with 1 as a substitute of 0.

The general approach to the decision tree algorithm we used, as summarized in Minnesota State University's data mining classes (cited by Bresfelean, 2009), was as follows:

1. We selected an attribute that best distinguishes the output attribute values.
2. Then, we generated a discrete tree branch for every value of the selected attribute.
3. The instances were partitioned into subgroups in order to reproduce the attribute values of the selected node.
4. For each subgroup, the attribute selection process was finished if:
    a. All members of a subgroup had the same value for the output attribute, the attribute selection process was finished for the present path and the branch on the present path was marked with the specified value.
    b. The subgroup included a sole node or no more individual attributes could be determined. As in (a), the branch was marked with the output value seen by the preponderant remaining instances.

5. For each subgroup generated in step (3) that has not been marked as terminal, the above process was.

In Weka (Witten et al., 2011), the algorithms' implementation are encapsulated in classes, that depend on other classes for some of their functionality. Every time the Java virtual machine executed J48 method, it generated an instance of this class by assigning memory for generating and storing a decision tree classifier. In the instantiation of the J48 class (Witten et al., 2011) were included the following: the algorithm, the classifier it built, and a procedure for outputting the classifier.

Bearing in mind the facts stated until now, after this thorough review of the research literature and due to the fact that we've identified little use of the method on the exact type of data that we would want to employ it on, we would like to proceed by further investigating the connections, given by data mining techniques, between raw data from different fields such as life standards, job, income, emigration, opinions and discontent, general and particular aspects of life and career, regarding our subjects life, collected thanks to the quality of life questionnaire.

4. Conclusions

The first part of the article allowed us to explore the concepts of this research: to identify and extract, from our main study, the key factors regarding one of the most worrying problems identified in the answers of our respondents – will to emigrate; to explore and stress out the importance of family and its unity during lifetime by discussing repercussions on its members; to analyze the problem of emigration from two points of view: personal and national/international; and set the grounds for the data processing method that we've used in identifying the possible motivations behind emigration decision by arguing its applicability and emphasis its novelty concerning the use of it on quantitative sociological data.

We consider that the findings of this research, which will be presented in the second part of this article, might be of interest by offering a plausible explanation to the motivation behind emigration decision, based on the income indicator, for Romanian teachers.

References

Amento B., Terveen L., Hill W., Hix D., and Schulman R. (2003). Experiments in social data mining: The TopicShop system. *ACM Trans. Comput.-Hum. Interact*, **10**(1), pp. 54-85.

ANBCC - Asociaţia Naţională a Birourilor de Consiliere pentru Cetăţeni (2005). *Românii si migraţia forţei de muncă în Uniunea Europeană*, Bucureşti.

Annoni P., Ferrari P.A., and Salini S. (2006). Data mining analysis on Italian family preferences and expenditures. In: Perner, P., ed. *Proceedings of the 6th Industrial Conference on Data Mining conference on Advances in Data Mining: applications in Medicine, Web Mining, Marketing, Image and Signal Mining (ICDM'06)*, Berlin, Heidelberg: Springer-Verlag, pp. 324-336.

Antman, F. (2013). The Impact of Migration on Family Left Behind. In: Constant, A. and Zimmermann, K. F., eds. *International Handbook on the Economics of Migration*. Northampton: Edward Elgar.

Bailey R. (2006). Physical Education and Sport in Schools: A Review of Benefits and Outcomes. *Journal of School Health*, **76**(8).

Berson A., Smith S. and Thearling K. (2000). *Building data mining applications for CRM.* McGraw Hill.

Bieber F.R., Brenner C. and Lazer D. (2005). Data mining the family tree: identification of relatives using genetic kinship analysis of DNA. In: *Proceedings of the 2005 national conference on Digital government research (dg.o '05)*. Digital Government Society of North America, pp. 239-240.

Botezat, A. and 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.

Breşfelean V. P. (2009). Data Mining Applications in Higher Education and Academic Intelligence Management. *In*: Meng, J., E. and Zhou, Y., ed. *Theory and Novel Applications of Machine Learning*, Vienna: I-Tech, pp. 209-228.

Breşfelean V. P. (20089. *Implicaţii ale tehnologiilor informatice asupra managementului instituţiilor universitare*, Cluj-Napoca: Ed. Risoprint.

CCRPL - Comisia Centrală pentru Recensământul Populaţiei şi Locuinţelor, *Comunicat de presă*, 2 februarie 2012, Bucureşti.

Chapman P., Clinton, C., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C. and Wirth, R. (2000). CRISP-DM 1.0, Step-by-step data mining guide, ftp://ftp.software.ibm.com/software/analytics/spss/support/Modeler/Documentation/14/UserManual/CRISP-DM.pdf (accessed August 12, 2014).

Chris, M., Wilson, A. and Oswald, J. (2005). *How Does Marriage Affect Physical and Psychological Health? A Survey of the Longitudinal Evidence*, IZA DP No. 1619, Bonn.

Chun A. H. W. (2007). Using AI for e-government automatic assessment of immigration application forms. *In*: Cheetham, W. ed. *Proceedings of the 19th national conference on Innovative applications of artificial intelligence - Volume 2 (IAAI'07)*, AAAI Press, pp. 1684-1691.

Ciuperca, N. (2009). *Efectele emigrării asupra familiei contemporane*, 1 May, http://ciupercaniculina.blogspot.ro/2009/05/efectele-emigrarii-asupra-familiei.html (accessed August 12, 2014).

Cuncea, C. (2012). *Peste 3,5 milioane de români au plecat din ţară în ultimii 40 de ani*, Gandul, 23 May 2012, http://www.gandul.info/financiar/studiu-in-ultimii-40-ani-3-5-milioane-de-romani-au-emigrat-romania-a-pierdut-40-50-mld-euro-9662507 (accessed August 12, 2014).

Eagle, N. and Pentland, A. (2005). *Reality mining: sensing complex social systems*, London: Springer-Verlag.

Fayyad, U., Piatetsky-Shapiro, G. and Smyth, P. (1996). *From Data Mining to Knowledge Discovery in Databases*, Menlo Park: American Association for Artificial Intelligence, pp. 37-54.

Filip F. (2000). Decizie asistată de calculator. Concepte, metode şi tehnici pentru deciziile centrate pe analiza datelor. *Revista Informatică Economică*, **4**(16).

Garcia, E., Romero, C., Ventura, S. and De Castro, C. (2009). An architecture for making recommendations to courseware authors using association rule mining and collaborative filtering. *User Modeling and User-Adapted Interaction*, **19**(1-2), pp. 99-132.

Hăisan, A.-A. (2013). *Disfuncționalități în Sistemul Educațional Național – studiu de caz – profesorii din învățământul preuniversitar clujean*. Cluj-Napoca: Presa Universitară Clujeană.

Ioniţă A. (2005). Asupra termenului de minerit de date (Data Mining). *Revista română de informatică şi automatică*, **2**, rria.ici.ro/ria2005_2/art03.html (accessed August 12, 2014)

Kotsiantis, S. B., Zaharakis, I. D. and Pintelas, P. E. (2006). Machine learning: a review of classification and combining techniques. *Artif. Intell. Rev.*, pp. 159-190.

Kotsiantis S. B. (2007). Supervised Machine Learning: A Review of Classification Techniques. *Informatica*, **31**, pp. 249-268.

Lin, C.-C. and Shih, D.-H. (2008). Associating Information Literacy with Regulating Rules in Family by Data Mining. In: *Proceedings of the 2008 3rd International Conference on Innovative Computing Information and Control (ICICIC '08)*. Washington: IEEE Computer Society, pp. 547.

Marbán, O., Mariscal, G. and Segovia J. (2009). A Data Mining & Knowledge Discovery Process Model. *In*: Ponce and Adem Karahoca, ed. *Data Mining and Knowledge Discovery in Real Life Applications*, Vienna: I-Tech, pp. 438-453.

Mikheenkova, M. A. (2011). Computer-Support Capabilities for Qualitative Research in Sociology. *Automatic Documentation and Mathematical Linguistics*, **45**(4), pp. 80–201.

Ogasawara, M., Sugimori, H., Iida, Y. and Yoshida, K. (20059. Analysis between lifestyle, family medical history and medical abnormalities using data mining method – association rule analysis. *In*: Rajiv Khosla, Robert J. Howlett and Lakhmi C. Jain, ed. *Proceedings of the 9th international conference on Knowledge-Based Intelligent Information and Engineering Systems - Volume Part II (KES'05)*, Berlin, Heidelberg: Springer-Verlag, pp. 61-171.

Marginean, I. et al. (2010). *Calitatea vieţii în România 2010*, Bucureşti: ICCV.

Mertens, P., Hagedorn, J., Fischer, M., Bissantz, N. and Haase, M. (1996). *Towards active management systems. Implementing Systems for Management Decisions; Concepts, Methods and Experience*. Chapman & Hall, pp. 305-325.

Mihai A. (2011). *2,1 milioane de români lucrează în 15 state europene*, Ziarul Financiar, 27 April 2011, http://www.zf.ro/analiza/2-1-milioane-de-romani-lucreaza-in-15-state-europene-8197333 (accessed August 12, 2014).

Moldoveanu, R. (2011). *Cei 10.000 de medici emigraţi vă salută de peste mări şi ţări*, Evenimentul Zilei, 19 November 2011, http://www.evz.ro/detalii/stiri/cei-10000-de-medici-emigrati-va-saluta-de-peste-mari-si-tari-954499.html (accessed August 12, 2014).

Nedelcu, A. et al. (2010). *Școala așa cum este*. Bucureşti: Centrul Educația 2000+, Unicef.

Pescaru, M. (2010). *Consecinţele migraţiei familiei contemporane asupra creşterii şi educării copiilor*, Cluj-Napoca.

Priţulescu, R. (2011). *Cum dispare populaţia României*, Adevărul, 13 June 2011, http://www.adevarul.ro/actualitate/Cum_dispare_populatia_Romaniei_0_497950541.html (accessed August 12, 2014).

Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann.

Romero, C., Ventura, S. and De Bra, P. (2005). Knowledge Discovery with Genetic Programming for Providing Feedback to Courseware Authors. *User Modeling and User-Adapted Interaction*, **14**(5), pp. 425-464.

Russell, M. A. (2011). *Mining the Social Web: Analyzing Data from Facebook, Twitter, Linkedin, and Other Social Media Sites*. 1st ed. O'Reilly Media, Inc.

Shah, S., Roy, R. and Tiwari, A. (2006). Technology Selection for Human Behaviour Modelling in Contact Centres. In: Rajkumar Roy and David Baxte ed. *Decision Engineering Report Series*, Cranfield University.

Stănculescu M. et al. (2011). *Impactul crizei economice asupra migraţiei forţei de muncă româneşti*, Bucureşti: Friedrich Ebert Stiftung.

Ullman, J. D. (2003). Data Mining Lecture Notes, Stanford University, http://infolab.stanford.edu/~ullman/ (accessed August 12, 2014).

Venkatadri, M, Loganatha, C. and Reddy (2011). A Review on Data mining from Past to the Future. *International Journal of Computer Applications*, **15**(7), pp. 19-22.

Warburton, D. et al. (2006). Health benefits of physical activity: the evidence, *CMAJ*, **174**(6), pp. 801-809.

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*, San Francisco, Morgan Kaufmann, pp. 702–707.

Witten, I. H., Frank, E. 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).