Global Journals LaTeX JournalKaleidoscope™
Artificial Intelligence formulated this projection for compatibility purposes from the original article published at Global Journals. However, this technology is currently in beta. *Therefore, kindly ignore odd layouts, missed formulae, text, tables, or figures.*

Association of Social Data

Diego C. Rodrigues[1], Marcelo Lisboa Rocha[2] and Jucelino Santos[3]

[1] IFTO-Federal Institute of Tocantins

Received: 16 December 2013 Accepted: 2 January 2014 Published: 15 January 2014

Abstract

According to the ILO (International Labour Organisation) about 218 million children between 5 and 17 years working in the world, of which 50

Index terms— data m ining, social data, child labor.

1 INTRODUCTION

ith the goal of minimizing social problems, the Brazilian Government created in October 24, 2001 the register only Brazilian (CadÚnico) **??**12], which is an instrument to record various information from low-income families, considering features of income up to $ 334 per person or per family monthly income of up to $ 1,002. Maintained for more than 13 years ago by the Government. This database has the function to log all Brazilian families who are under social risk. The CadÚnico stores a set of data on households and their members, creating a powerful set of data with great potential for information.

CadÚnico's database allows understanding the socioeconomic reality of these families, bringing information on the nuclear family, such as the characteristics of their residences, the forms of access to essential public services and, also, the information about each of the components of the family [11].

The Federal Government, through a computerised system consolidates data collected at CadÚnico. With this information in hand, the Government can use them to formulate and implement specific policies that contribute to the reduction of social vulnerabilities that these families are exposed [13].Currently, the program has more than 21 million households registered in Brazil, being coordinated by the Ministry of Social development and fight against Hunger (MDS) and should be used for selection of beneficiaries of social programs of the Federal Government [13].

Applying data mining techniques to discover patterns and valid knowledge is not an easy task, due to the large amount of data and attributes available on CadÚnico. With the purpose of analyzing these data, applied knowledge discovery techniques to find standards regarding child labour factors in Tocantins State.

According to the IBGE (Brazilian Institute of geography and statistics), in the records of the 2010 census, the northern region of Brazil is in first place on indexes of child labour. From this, some questions need to be identified, and the need to highlight the situation of these families pointed out by IBGE, in order to ensure that the actions of the Government have the capacity to achieve and support these people [10].

This work is intended to answer the following questions: "what are the deterministic factors which prove the non-existence of the practice of child labour?" and "what are the real relationships between the regional indicators, financial, educational, cultural and social sciences?".In fact, there is a strong tendency to associate the child labour to family financial condition of the child, but this factor is actually a deterministic indicator? And when you consider other indicators? Therefore, this paper aims, through the application of data mining techniques, find in CadÚnico, patterns that indicate with a high degree of relevance to nonexistence of the practice of child labour.

The remainder of the work is organized as follows. Section II describes the Organization of data, section III describes the process of knowledge discovery used, section IV contextualizes the Association Rules algorithm, methodology, section V section VI section VII, experiments, results VII conclusion.

2 II.

3 ORGANIZATION OF DATA

The database is divided into two groups, being they family and individuals. In the set of data on families are all information about type of housing, family income and social information, in addition to the financial and regional conditions.The data about the individuals belonging to the families are their characteristics and personal data, such as school, social and financial information for each individual. Thus, the Government generates a complete record of families and their members collecting these data every 3 months over the course of a year so there's 4 data collections each year.

For this work were used initially 92 database attributes, 5 attributes were selected through the W

4 THE PROCESS OF KNOWLEDGE DISCOVERY

Data mining is part of a larger research process, known as knowledge discovery in databases, being defined as the exploration and analysis, automatic or semiautomatic, large amounts of data to find patterns and relevant rules that are not easily found/obtained [1].

Data mining techniques are growing in popularity as a tool for knowledge discovery in the search for unknown information, necessary for the decision-making process. However, this approach is difficult to apply because of their interdisciplinary skills to combine different methods and techniques, such as database, statistical methods, neural networks, genetic algorithms, machine learning, natural language processing and other fields of study.

The data to be worked are not always in perfect condition to start the process of mining. Data usually have many sources, may be incomplete or contain noise. Data needs to be treated by an important step, known as preprocessing, including activities such as cleaning, integration, selection and processing of data [2].

After the completion of preprocessing, data can be loaded into the data mining software, such as, for example, the WEKA (Waikato Environment for Knowledge Analysis) [3]. The WEKA includes a number of algorithms for formatting data, algorithms for machine learning and validation of results, being written in the Java programming language [4], having its open source and available on the internet.

One of the data mining techniques available, the task of association can be considered one of the most appropriate techniques for the purpose of possible applications for these rules. The goal of the analysis of the data was generate Association rules regarding social indicators in order to discover patterns that weren't explicit about the data related to child labour. Further details about membership rules and its algorithm of production will be described in section IV below.

IV.

5 Rules of Association

The Apriori algorithm was proposed by r. Agrawal and r. Srikant in 1994 for mining frequent items series in databases.The name of the algorithm is based on the fact that your method using the characteristics of a frequent pattern already found earlier (prior) to get some more patterns [5].

Association rules have as their basic premise find elements that imply in the presence of other elements in the same transaction, i.e. find frequent patterns or relationships between datasets.The term transaction indicating which items have been found at a particular query operation. Various metrics can be used to evaluate the rules and identify which are interesting. The most commonly used restrictions are supported and trusted.

The Association rule has the form A B, where A, called the antecedent, and B, called the consequent, are sets of items or transactions, and the rule can be read as: often attribute attribute implies B (Agrawal; Imielinski; Swami, 1993) **??**14].

To evaluate the rules generated some interest measures are used, the most used are support and confidence (described in paragraph below) also scientific studies. The APRIORI algorithm was chosen established on work (Agrawal; Imielinski; Swami, 1993).The authors (Geng and Hamilton, 2006) carried out a survey involving other metrics and suggested for generating Association rules with a wide range of strategies for selecting appropriate measures for certain domains and requirements **??**14]. In this work we use the following measures:

? Support : P (AB). The support of a rule is defined to be the fraction of items that meet the set A and B of the rule. If support is not large enough, this means that the rule is not worthy of consideration or that is simply deprecated and can be considered later [14]; ? Confidence: P (A/B). Is a measure of the strength of the support rules and corresponds to statistical significance. The likelihood of finding the part B of rule in transactions on the condition that these transactions also contain the [14]. ? Interest (lift): P (B |The)/P (B) or P (AB)/P (A) P (B).

Used to find dependencies, it indicates how much more often becomes B when occurs. Varies between 0 and ? [14].

V.

6 AS EVERYTHING WAS DONE

The first stage of the work consisted of examining data mining algorithms and choose the most suitable for finding patterns between the attributes in analysis. The APRIORI algorithm was appointed by the IEEE International

Conference on Data Mining (ICDM) [6] as the most promising algorithm for association rules generation and one of the most popular approaches in data mining. Therefore, it was considered in the realization of this work.

7 a) Preprocessing

The only data quality are considered, provided they satisfy the requirements of the intended use. There are many factors that make up the quality of the data, including the accuracy, completeness, consistency, timeliness, credibility and interpretability [7]. To ensure these measures of quality in preprocessing, some steps are needed, as follows.

8 b) Integration of Data

The CadÚnico database was stored in different tables. To assist in reducing redundancies and inconsistencies in the data set was held a secure integration, which used the sample code, unique key link between the table that contained information of persons and of their families. The final set was then raised in a CSV file (comma-separated values). Redundant data were grouped or eliminated depending on the value in relation to the sample, avoiding inconsistencies in the data set.

9 c) Data Cleaning

At this stage some routines were performed in an attempt to ensure the quality of the data, where d) Data Reduction After cleaning the data, the final set of attributes has been reduced from the original, performing a downsizing where weakly relevant attributes or redundant data might be detected and eliminated.

In this task he was employed the CfsSubsetEval algorithm to evaluate the value of a subset of attributes, whereas the predictive power of each feature, as well as the degree of redundancy between them.The subsets of features highly correlated with class and with low intercorrelação are preferably selected [7].

Para este trabalho, a combinação entre BestFirst (método de busca) e CfsSubsetEval (atributo avaliador) é tão eficiente quanto as técnicas de seleção de variáveis, como algoritmo genético e o algoritmo SimulatedAnnealing, além de ser mais rápido [6].

To evaluate the attributes, values were compared using the heuristics of the merit of each relationship formalized by the equation of the formula I. The final formula of merit using the Pearson correlation coefficient between a composite variable (sum or average) and a target variable (the class in question) [6].

The CfsSubsetEval algorithm implemented in WEKA was executed with the initial set of data as input. Of 92 initial attributes, the base was reduced to 35.After the assessment of experts in social work this number was reduced again and totaled 5 attributes considered essential for modeling of the problem. The list of attributes with their respective merit scores is presented in Figure **??**.

10 Figure 1 : Result CfsSubsetEval

In addition, the data were related to each family with their respective members and then, after running the above attributes selection algorithm, the same attributes selected by the algorithm were elected by the experts as the most relevant indicators for the problem in question.

11 a) Data Transformation

The tabs "," decimal numeric values were replaced with "." for correct interpretation by WEKA and null values were replaced with "?", and processed numeric data for nominal.

VI.

12 COMPUTATIONAL EXPERIMENTS

The first step in the phase of experiments was the identification of the non-conformity of data quality especially for the amount of missing data in the CadÚnico database.This evidence requires a refinement of the data in 888,621 records, verifying the existence of blank fields and null values for the 6 selected attributes (sex of child, family income, school attendance, child labour rate, existence of federal program assistance and region of the State that the city is located), in order to adjust the database to be interpreted by the WEKA software package.

In the second stage was held the selection and data junction with the language use Structured Query Language (SQL), which supports data manipulation through the selection of records without missing values in its fields. This step allowed validate and prepare all data to be exported to the CSV file format.

The third stage consisted in the identification and collation of the most relevant indicator of child labour. The goal was to create a set of data from the CadÚnico base representing only the records in which the attribute "child labor" was marked, positively or negatively, leaving out all the records that have this field with missing values.The result was a total of 300,415 records with this field filled in. This step allowed to evaluate the context of data reliably and apply the techniques of knowledge discovery.

In the fourth step, the database is separated into two parts, with a 70% of data, were considered for training and the other with 30% of the data were used for testing.The separation of these two parts was conducted

randomly, with the use of the SQL functions, with the goal of having a more efficient learning (no trend) in the use of the algorithms in WEKA.

Finally in the fifth step was performed to validate the refinement of CadÚnico database with experts in government social assistance.All selected indicators were presented to the staff of the Ministry of Social development and hunger alleviation so that the dataset was evaluated and validated by them.Thus, he made sure the dataset was reliable and could be used in data mining. In the end, the quality of the generation of association rules were evaluated by a group of specialists in social assistance, the Secretariat servers work and Social Assistance of Tocantins, where it was evaluated the classification of the indicators as their relevance in relation to the possible factors of the existence or non-existence of child labour in relation to the attributes worked.

Upon completion of the preprocessing step, the data set was imported by WEKA for data to be used in the APRIORI algorithm, generating Association rules.

13 VII.

14 RESULTS OBTAINED

From the implementation of APRIORI algorithm, 35 rules were obtained with a confidence of 90% and minimum support of 10%. All the rules generated were presented to an expert of human and social areas of the Secretariat of social welfare of the State of Tocantins, where they were separated and validated.

Based on the knowledge acquired was created a map as shown in Figure **??** The formation of the map is composed of all occurrences of child labour identified, separated by region with the use of the ind_trabalho_infantil_pessoa region attributes as shown in the session II, applying the techniques presented in the session SAW and with a software able to relate the occurrences with the geographic coordinates (Latitude and Longitude) to create the view proposed in Figure 2.

For better understanding, the rules generated were organized in the most relevant attributes and interpreted according to In Figure **??**, we have a rule that says with 98% confidence that the database used, when there is Government assistance for this family, the index of child labour is non-existent, on the basis of this rule was created the map in Figure 2. Soon, when you know when you don't have the factors that notify when child labour occurs one can easily find out when they occur, thanks to good faith rule. Observing the schooling and income factor in Figure 5, it may be noted that even if income finding low, factors like education are necessary to eradicate child labour, where school attendance has been confirmed, the indexes were summarized as shown in Figure 2 in the central region of the map, thus confirming the rule of Figure **??**, where who receives Government assistance does not exist cases of child labour. Soon, to be entitled to this benefit, families have to comply with the rules established by the Brazilian Government that deals with the requirement for school children.

All the rules used to generate the map in Figure 2 were applied with a minimum of 30% with a minimum of 90% confidence, therefore, the set of rules presented covers the assertions here presented, thus giving support to managers in the decision-making process, with data collected by the federal Government of Brazil.

15 VIII.

16 CONCLUSION

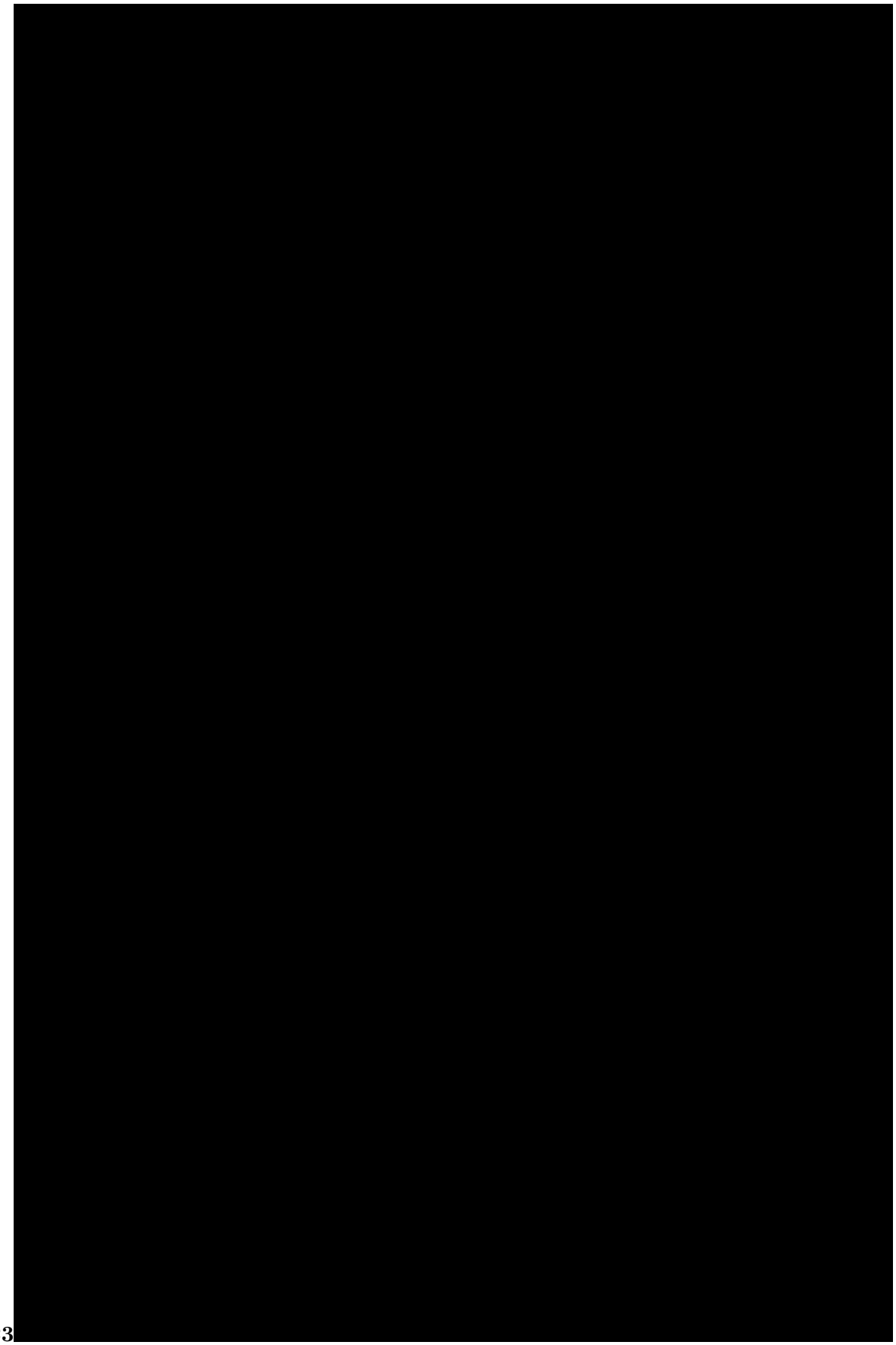
In Brazil, policies to combat child labour are offered by the Federal Government. Apply knowledge discovery techniques and patterns is a way to leverage the results to be analyzed, aiding the identification, location and better understanding of the cases of children who are in situations of social risk, especially in Brazil which is the 5° largest country in geographic extent of the world.Data mining can allow effective searches for potential risk social activities. In this way, the knowledge discovery can provide decision makers with information and knowledge of what actions are required to combat the various factors of this problem that is global.

Using computational techniques shown in section VI, it was possible to identify what are the deterministic factors which prove the non-existence of the practice of child labour, based on the knowledge discovered may also identify the cases that occur child labor as was demonstrated on a map of Figure 2 and relating between indicators and rules generated, through the techniques of knowledge discovery in section III, we observed a strong relationship between child labor with regional factors, school attendance and Government assistance.

The results of this survey suggest how further work studies that can perform the comparison of international data as well as the development of software that support managers in decision-making and development of a computational modeling to serve as a tool for everyone who need to solve the same type of problem.

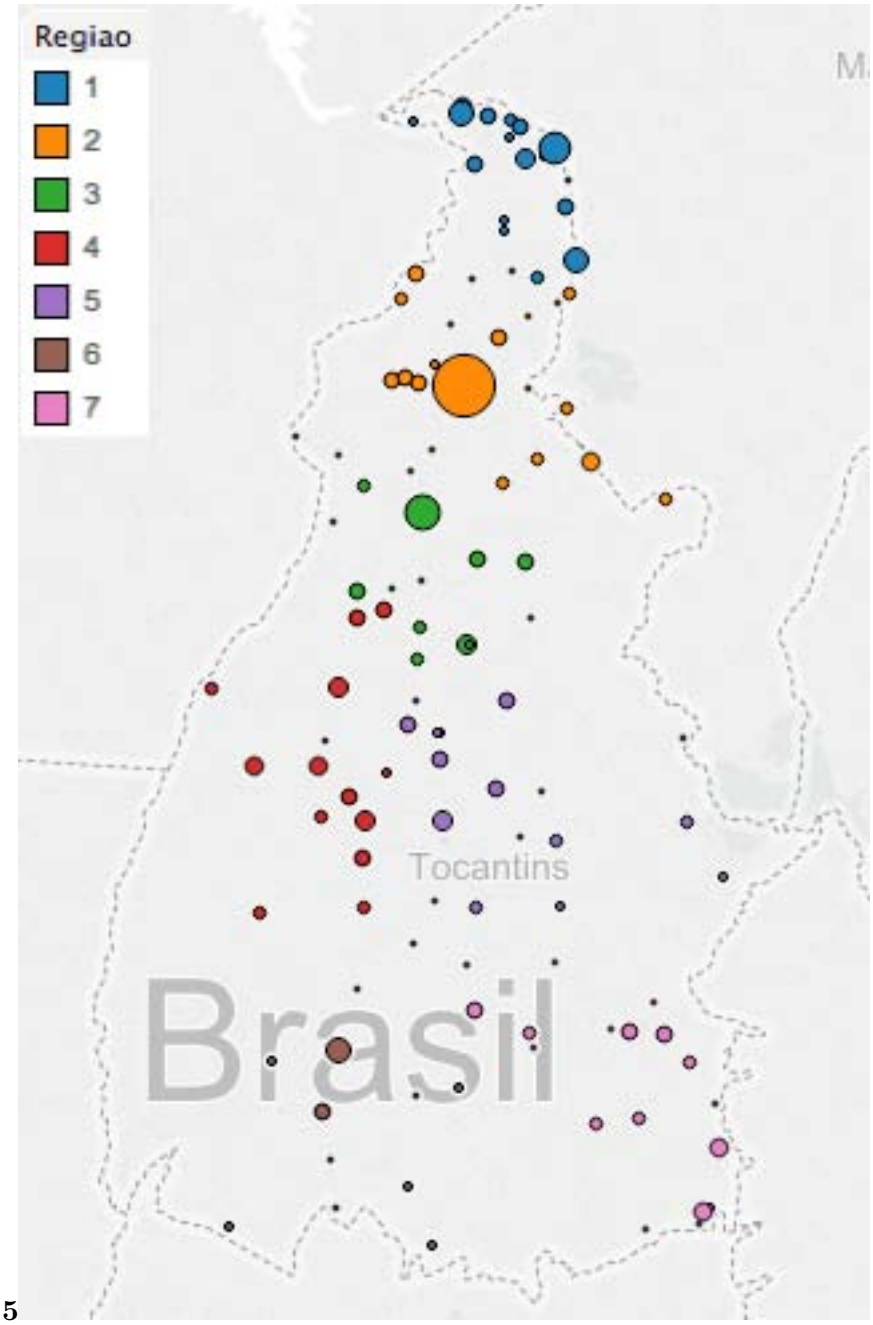
2

Figure 1: Figure 2 :



23

Figure 2: Figure 2 .Figure 3 :



5

Figure 3: Figure 5 :

1

Attributes	Possible Values		Meaning
Regiao	Territorial division of the State of	?	1-Norte I
	Tocantins:	?	2-North II
	1,2,3,4,5,6,7	?	3-North III
		?	4-Midwest
		?	5-East Center
		?	6-Southwest
		?	7-Southeast
marc_pbf	If Receives Government	?	0-No
	Assistance:0, 1	?	1-Yes
ind_frequenta_escola_memb	School attendance of the	?	1-Yes, public network
	person:1, 2, 3, 4	?	2-Yes,private network
		?	3-no, I attended
		?	4-Never attended
fx_rfpc	Family Income:1, 2, 3, 4	?	1-$ 28.00
		?	2-from $ 28.00 to $ 56.00
		?	3-from $ 56.00 to $ 140.00
		?	4-US $ 140.00
ind_trabalho_infantil_pessoa	Person with index of child	?	1-Yes
	labour:1,2	?	2-No

III.

Figure 4: Table 1 :

Attribute	Merit	Selected Attributes
ind_trabalho_infantil_pessoa	0.133	Regiao
	?	cod_sexo_pessoa
	?	ind_frequenta_escola_memb
fx_rfpc	0.114	? Regiao
	?	marc_pbf
ind_frequenta_escola_memb	0.113	marc_pbf
	?	ind_trabalho_infantil_pessoa
marc_pbf	0.114	? ind_frequenta_escola_memb
	?	fx_rfpc
Regiao	0.016	? fx_rfpc
	?	ind_trabalho_infantil_pessoa

geographical location of all known cases of child labour in the State in the year 2014 separated by regions.

Figure 5:

Relationship Rules and attributes	Confidence
6. Regiao=5 marc_pbf=1 ==> ind_trabalho_infantil_pessoa=2	conf:(0.99)
21. Regiao=1 marc_pbf=1 ==> ind_trabalho_infantil_pessoa=2	conf:(0.98)
22. Regiao=2 marc_pbf=1 ==> ind_trabalho_infantil_pessoa=2)	conf:(0.98)

Figure 4 : Rules to generate region 5,1 and 2 in map

Relationship Rules and attributes	Confidence
27.ind_frequenta_escola_memb=1fx_rfpc=2==>ind_trabalho_infantil_pessoa=2	conf:(0.98)
28. ind_frequenta_escola_memb=1 ==> ind_trabalho_infantil_pessoa=2	conf:(0.98)
29.ind_frequenta_escola_memb=1fx_rfpc=1==>ind_trabalho_infantil_pessoa=2	conf:(0.98)
32.ind_frequenta_escola_memb=1fx_rfpc=3==>ind_trabalho_infantil_pessoa=2	conf:(0.98)

Figure 6:

[Rodrigues et al.] , D C Rodrigues , D N Prada , M A Silva .

[Hall et al. ()] , Mark Hall , Eibe Frank , Geoffrey Holmes , Bernhard Pfahringer , Peter Reutemann , Ian H Witten . 2009.

[Ibge ()] , Ibge . http://censo2010.ibge.gov.br/trabalhoinfantil August 2013. 2010.

[)law creating the Brazilian single register (2014)] *)law creating the Brazilian single register*, http://www.planalto.gov.br/ccivil_03/_ato2004-2006/2004/lei/l10.836.htm November 2014. Planalto. 12.

[Igor et al. ()] 'Benchmarking of Linear and Nonlinear Approaches for Quantitative Structure ?Property Relationship Studies of Metal Complexation with Ionophores'. V Igor , ? Tetko , P Vitaly , Alexey V Solov'ev , Xiaojun Antonov , Jean Pierre Yao , Botao Doucet , Frank Fan , Denis Hoonakker , Piere Fourches , Nicolas Jost , Alexandre Lachiche , Varnek . *Journal of Chemical Information and Modeling* 2006. 46 (2) p. .

[Hall ()] *Correlation-based Feature Selection for Machine Learning*, M A Hall . 1998. Hamilton, NZ. Waikato University (Ph.D diss)

[Linoff and Berry ()] *Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management*, Gordon S Linoff , Michael J Berry . 2011. Wiley. (3rd ed.)

[Prata ()] 'Exploring an Ichthyoplankton Database from a Freshwater Reservoir in Legal Amazon'. Monica Prata . Lecture Notes in Computer Science. 1ed.Hangzhou 2013. Springer. p. .

[Exploring Social Data to Understand Child Labor International Journal of Social Science and Humanity ()] 'Exploring Social Data to Understand Child Labor'. *International Journal of Social Science and Humanity* 2015. (5) p. .

[Agarwal et al. (1996)] 'On the computation of multidimensional aggregates'. S Agarwal , R Agrawal , P M Deshpande , A Gupta , J F Naughton , R Ramakrishnan , S Sarawagi . *Proc. 1996 Int. Conf. Very Large Data Bases (VLDB'96)*, (1996 Int. Conf. Very Large Data Bases (VLDB'96)Bombay, India) Sept. 1996. p. .

[Operation of registration only Brazilian MDS (2014)] 'Operation of registration only Brazilian'. http://www.mds.gov.br/bolsafamilia/cadastrounico *MDS* November 2014.

[References Références Referencias] *References Références Referencias*,

[SIGKDD Explorations] *SIGKDD Explorations*, 11.

[The Java Programming Language and the Java Platform (2013)] *The Java Programming Language and the Java Platform*, http://www.oracle.com/technetwork/topics/newtojava/downloads/index.html August 2013.

[The WEKA Data Mining Software: An Update] *The WEKA Data Mining Software: An Update*,

[Wu et al. (2007)] 'Top 10 algorithms in data mining'. Xindong Wu , Vipin Kumar , J Ross Quinlan , Joydeep Ghosh , Qiang Yang , Hiroshi Motoda , Geoffrey J Mclachlan , Angus Ng , Bing Liu , Philip S Yu , Zhi-Hua Zhou , Michael Steinbach , David J Hand , Dan Steinberg . *Knowl. Inf. Syst* 2007. December 2007. 14 p. .

[O Estadao De et al. ()] *Worldwide, 218 million children work*, S O Estadao De , E E Paulo , Ghiselli . http://www.estadao.com.br/noticias/internacional,218-milhoes-de-criancas-trabalham-no-mundo-calcula-oit August 2013. 10222,0.htm 9. 1964. 1964. McGraw-Hill. (Theory of psychological measurement)