RETIREMENT HOME SMOKING CESSATION [1] [2]

João Emmanuel D'Alkmin Neves[3]
Antônio André Cunha da Silva[4]

ABSTRACT

This paper analyzes the opinion of health staff members about retirement home residents on tobacco cessation through a data mining approach. This analysis mainly aims the identification of which type of professional is more willing to give advice to people who live in retirement home. In this manner, a bunch of questions were answered by the health staff and converted into a amount of variables. These variables were preprocessed and cleaned and used as entry to sophisticated methods using the Data Mining Program called Waikato Environment for Knowledge Analysis (WEKA). Those methods were classification, association, and clustering. As a result, we had the most willing profession.
Keywords: Data mining ; Retirement Home ; Smoking Cessation

RESUMO

Este artigo analisa a opinião de profissionais da área da saúde sobre residentes de asilo e o abandono do vício do cigarro através de uma abordagem de mineração de dados. Esta analise foca na identificação de qual tipo de profissional é mais provável a dar conselhos a favor e contra sobre o abandono do cigarro em pessoas idosas que vivem em asilos. Desta forma, uma grande quantidade de perguntas foi respondida pelos profissionais e convertidas em um conjunto de variáveis. Estas variáveis foram pré-processadas e filtradas, e em seguida, usadas como entrada de sofisticados métodos executados com o apoio do programa de mineração de dados (WEKA). Estes métodos são chamados de classificação, associação e agrupamento.
Palavras-chave: Mineração de dados ; Asilo ; Tabagismo

INTRODUCTION

People that have different jobs in health science have answered a set of questions about advising for retirement home residents on tobacco cessation. These questions were answered by professionals of health area. The main objective of this project is to find relationships between the profession of the health staff and their attitudes and beliefs about retirement home residents on tobacco cessation. This becomes a very interesting topic because even though many of them known that tobacco is very harmful to the health of people, in this case, retiree have a short period of life. Therefore, there is big discussion about if it is worth for these people to quit smoking and to suffer all the side effects left for the abstinence. The work for this analysis was separated in four important phases in data mining analysis, they are:data preprocessing, classification, association, and clustering. It is important to mention that an important variable called position is analyzed and processed during all the project.

1 STATISTICAL ANALYSIS AND DATA PREPROCESSING

Statistical analysis of data is the activity that aims to analyze and to transform a set of data in order to improve the verification and identification. According to Gibilisco (2011), statistical analysis is a component of data analytics. In the context of business intelligence (BI), statistical analysis involves collecting and scrutinizing every single data sample in a set of items from which samples can be drawn. Statistical analysis can be broken down into five discrete steps, as follows: describe the nature of the data to be analyzed; explore the relation of the data to the underlying population; create a model to summarize understanding of how the data relates to the underlying population; prove (or disprove) the validity of the model; and employ predictive analytics to run scenarios that will help guide future actions.

The goal of statistical analysis is to identify trends. A retail business, for example, might use statistical analysis to find patterns in unstructured and semi-structured customer data that can be used to create a more positive customer experience and increase sales. Statistical analysis of data has different methods and

[1] Artigo desenvolvido como aplicação do uso de data mining em um estudo de caso, baseado em dados reais, disponibilizados pelas próprias instituições norte-americanas. - Disciplina: Data Mining - Professor: Dr. Anthony Scime - Instituição: State University of New York
[2] Enviado para submissão em 18/07/2015
[3] Tecnólogo em Tecnologia em Análise e Desenvolvimento de Sistemas – Fatec Americana – Centro Estadual de Educação Tecnológica Paula Souza ; Contato: :jeneves@gmail.com
[4] Pesquisador da Universidade Federal do Pará.

approaches using various techniques. Data can be of several types, such as quantitative data, categorical data, and qualitative data.

After the statistical analysis of data is started data preprocessing step. This process comprises the application of various techniques for collecting, organizing, processing, and preparing the data. According to Garcia, Luengo and Herrera (2014), data preprocessing includes data preparation, compounded by integration, cleaning, normalization and transformation of data; and data reduction tasks; such as feature selection, instance selection, discretization, etc. The resulted expected after a reliable chaining of data preprocessing tasks is a final dataset, which can be considered correct and useful for further data mining algorithms. It is a step that has fundamental relevance that extends from erroneous data correction by adjusting the formatting of data for data mining algorithms to be used.

To determine position of respondent about retirement home smoking cessation, numerous data mining processes are applied. Previously there are the preprocessing the data. This preprocessing was necessary for binning the data later in the analysis.

The main functions of data preprocessing used in the current project:

- Selection of attributes: aims to choose a subset of attributes or create other attributes that replace one set to reduce the size of the database. With this size reduction, it reduces the complexity of the database and thus the processing time for extracting some knowledge of it. In addition, attributes may cause unnecessary noise on the final result and it can be avoided by applying feature selection techniques.
- Data cleaning - covers any treatment performed on the selected data to ensure the quality, completeness, accuracy, and integrity of the facts they represent. Missing information, erroneous or inconsistent in the databases should be corrected or removed so as not to compromise the quality of the models of knowledge to be extracted at the end of the process.

The database used initially had 62 attributes and 647 records. The first step was the removal of null values. All data containing null values are removed from the database to avoid noise and unwanted results. The Microsoft Excel was the tool utilized to perform this activity. In total, 1163 data were removed in this first stage of the project.

Figure 1 Dataset with null values

q10a	q10b	q10c	q10d	q10e	q10f	q10g	q10h	q10i	q10j
2.0	2.0	1.0	2.0	4.0	3.0	2.0	4.0	2.0	4.0
4.0	3.0	4.0	5.0	4.0	4.0	3.0	4.0	2.0	4.0
3.0	2.0	2.0	5.0	2.0	5.0	1.0	4.0	3.0	3.0
2.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	3.0	5.0
4.0	3.0	1.0	1.0	1.0	5.0	5.0	5.0	1.0	1.0
2.0	5.0	5.0	5.0	5.0	2.0	4.0	4.0	4.0	#NULL!
4.0	4.0	4.0	5.0	5.0	4.0	2.0	2.0	1.0	4.0
3.0	3.0	4.0	5.0	3.0	3.0	5.0	4.0	2.0	3.0
#NULL!	4.0	5.0	5.0	4.0	3.0	4.0	2.0	2.0	2.0
4.0	4.0	3.0	3.0	4.0	5.0	4.0	4.0	3.0	4.0
3.0	4.0	3.0	3.0	4.0	5.0	5.0	3.0	1.0	4.0
3.0	3.0	#NULL!	1.0	1.0	4.0	3.0	4.0	2.0	3.0
3.0	3.0	2.0	2.0	3.0	4.0	3.0	4.0	2.0	4.0
2.0	5.0	5.0	5.0	5.0	2.0	1.0	1.0	1.0	1.0
2.0	4.0	4.0	4.0	2.0	4.0	2.0	1.0	1.0	2.0
2.0	4.0	4.0	4.0	4.0	2.0	4.0	4.0	2.0	4.0
1.0	#NULL!	2.0	1.0	4.0	3.0	1.0	3.0	5.0	2.0
4.0	4.0	4.0	5.0	4.0	4.0	4.0	4.0	4.0	4.0
1.0	1.0	4.0	#NULL!	1.0	1.0	1.0	1.0	4.0	4.0
1.0	1.0	3.0	1.0	#NULL!	3.0	1.0	3.0	#NULL!	1.0
#NULL!	1.0	4.0	4.0	4.0	4.0	1.0	1.0	1.0	1.0

Source: Author

Figure 2 Dataset without null values

q10a	q10b	q10c	q10d	q10e	q10f	q10g	q10h	q10i	q10j
2.0	2.0	1.0	2.0	4.0	3.0	2.0	4.0	2.0	4.0
4.0	3.0	4.0	5.0	4.0	4.0	3.0	4.0	2.0	4.0
3.0	2.0	2.0	5.0	2.0	5.0	1.0	4.0	3.0	3.0
2.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	3.0	5.0
4.0	3.0	1.0	1.0	1.0	5.0	5.0	5.0	1.0	1.0
2.0	5.0	5.0	5.0	5.0	2.0	4.0	4.0	4.0	
4.0	4.0	4.0	5.0	5.0	4.0	2.0	2.0	1.0	4.0
3.0	3.0	4.0	5.0	3.0	3.0	5.0	4.0	2.0	3.0
	4.0	5.0	5.0	4.0	3.0	4.0	2.0	2.0	2.0
4.0	4.0	3.0	3.0	4.0	5.0	4.0	4.0	3.0	4.0
3.0	4.0	3.0	3.0	4.0	5.0	5.0	3.0	1.0	4.0
3.0	3.0		1.0	1.0	4.0	3.0	4.0	2.0	3.0
3.0	3.0	2.0	2.0	3.0	4.0	3.0	4.0	2.0	4.0
2.0	5.0	5.0	5.0	5.0	2.0	1.0	1.0	1.0	1.0
2.0	4.0	4.0	4.0	2.0	4.0	2.0	1.0	1.0	2.0
2.0	4.0	4.0	4.0	4.0	2.0	4.0	4.0	2.0	4.0
1.0		2.0	1.0	4.0	3.0	1.0	3.0	5.0	2.0
4.0	4.0	4.0	5.0	4.0	4.0	4.0	4.0	4.0	4.0
1.0	1.0	4.0		1.0	1.0	1.0	1.0	4.0	4.0
1.0	1.0	3.0	1.0		3.0	1.0	3.0		1.0
	1.0	4.0	4.0	4.0	4.0	1.0	1.0	1.0	1.0

Source: Author

The second step was a thorough analysis of the codebook to identify which attributes were important to the project objective. This stage began with the analysis, understanding, and identification of related attributes. Due to the codebook, present incomplete information was necessary to conduct intensive research to understand the meaning of the attributes in order to determine which attributes were the determinants for the project. The attributes that did not present correlation with the main objective were removed, such as comments, license collapsed, and barriers. At the end of the second stage, all entries present in the codebook were translated and 46 attributes. These attributes were considered strongly related to the main theme of the project.

Figure 3 Codebook after second phase

Attribute	Type			Meaning
q10a	Numeric	8	1	important problem than smoking
q10b	Numeric	8	1	importance health benefits of cessation
q10c	Numeric	8	1	harmful to health
q10d	Numeric	8	1	second-hand smoke harmful
q10e	Numeric	8	1	quitting smoking improve the health
q10f	Numeric	8	1	right to smoke
q10g	Numeric	8	1	assisting residents is staff duty
q10h	Numeric	8	1	good for socialization
q10i	Numeric	8	1	have little to live for
q10j	Numeric	8	1	few pelasures residents have
q10k	Numeric	8	1	not capable of making decisions
q10l	Numeric	8	1	it is addiction
q10m	Numeric	8	1	it is a dirty habit
q10n	Numeric	8	1	adequate fire and safety precautions
q10o	Numeric	8	1	All located in a separate wing
q10p	Numeric	8	1	Require more care
q10q	Numeric	8	1	reinforcements for desired behavior
q10r	Numeric	8	1	problem list
q10s	Numeric	8	1	unsafe situations

Source: Author

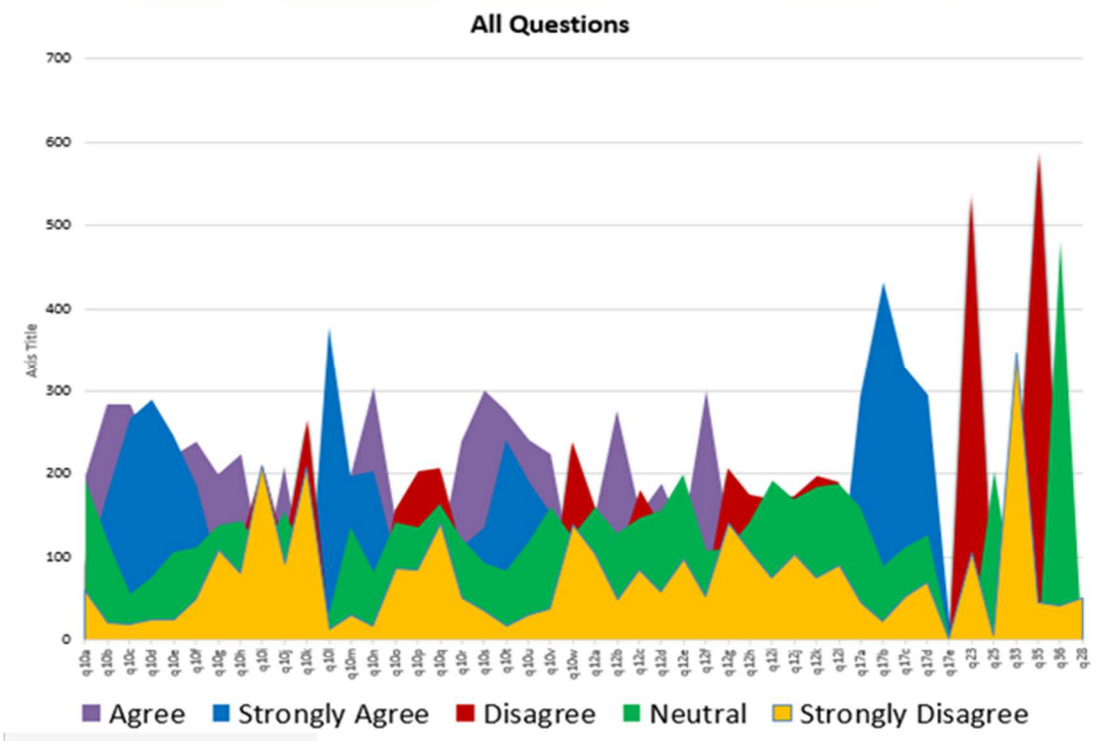
The third phase consisted of grouping the data of the attributes in a frequency distribution to show the quantities of the responses of professionals who answered Strongly Agree, Agree, Neutral, Disagree or Strongly Disagree with the issues presented. From this frequency distribution was generated a graph to a better view of the positions of professionals regarding the issues raised.

Figure 4 Frequency Distribution

	A	B	C	D	E	F
1		Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
2	q10a	59	106	192	198	84
3	q10b	21	36	120	284	176
4	q10c	18	18	55	283	267
5	q10d	25	26	76	225	290
6	q10e	25	31	105	222	247
7	q10f	49	55	111	239	186
8	q10g	108	109	138	200	87
9	q10h	80	118	143	223	68
10	q10i	209	209	108	80	32
11	q10j	91	112	154	208	68
12	q10k	207	265	98	49	21
13	q10l	13	17	31	201	377
14	q10m	30	79	135	198	197
15	q10n	16	29	81	305	204
16	q10o	86	157	141	149	101

Source: Author

Figure 5 Graphic Distribution



Source: Author

2 CLASSIFICATION MINING

Following the Statistical Analysis and Data Preprocessing phase, it was started the Classification Mining phase.

Classification mining is the step most used among the various data mining tasks. The classification consists in the discovery of forecasting rules to aid in planning and decision-making. In accordance with Kumar (2010), the classification mining is used when there are many records in a database with various attributes and it is necessary to extract relevant knowledge with predictive ability. This phase can be a alternative or complementary conclusion to the Association Phase.

In agreement with Witten, Frank and Hall (2011), in order to a better understanding of this technique it is necessary to understand some fundamental properties, such as comprehensibility and validity. The comprehensibility of the discovered knowledge is important because it will be used for decision-making. To facilitate understanding of the discovered knowledge is used knowledge representation by means of prediction rules. In this context, it is important to know the meaning of related attributes and main attribute of a rule. The ultimate goal of ranking algorithm is to generate the type rules: "If / Then". In this type of rule antecedent may contain one or more attributes, each associated with a value, and the resulting target must have the attribute. In addition to comprehensibility, another important property is the validity of the discovered knowledge. Each generated rule has a hit rate. The hit rate indicates the validity of a rule.

First, the objective question had to change its values from continuous values to descriptive values. It was necessary to change the values in order to prepare them for the WEKA software.

Figure 6 Discriptive Values

EducationLevel	Position	SmokingStatus
9	Doctor1	1
3	Doctor8	1
3	Doctor8	1
5	Doctor6	1
3	Doctor8	4
9	Doctor1	1
4	Doctor7	2
3	Doctor8	1
4	Doctor7	1
5	Doctor9	4
3	Doctor8	4
4	Doctor7	1
5	Doctor5	1
3	Doctor8	1
9	Doctor2	1

Source: Author

Second, the data set were divided into two sets: one set for training and testing set. The training set had 430 instances and it consists of data where the algorithm will find the classification rules. The test set had 216 instances and it is used to verificating of the discovered rules through the training set, to calculate the hit rate. Thus, the test suite acts as if their data were still unknown, as if they only appear in the future. Thus, it was possible to verify the validity of the previously defined forecasts.

Subsequently running the training set monitored by the test set on WEKA software, increasing the confidence value each time for comparison purposes, these results were obtained:

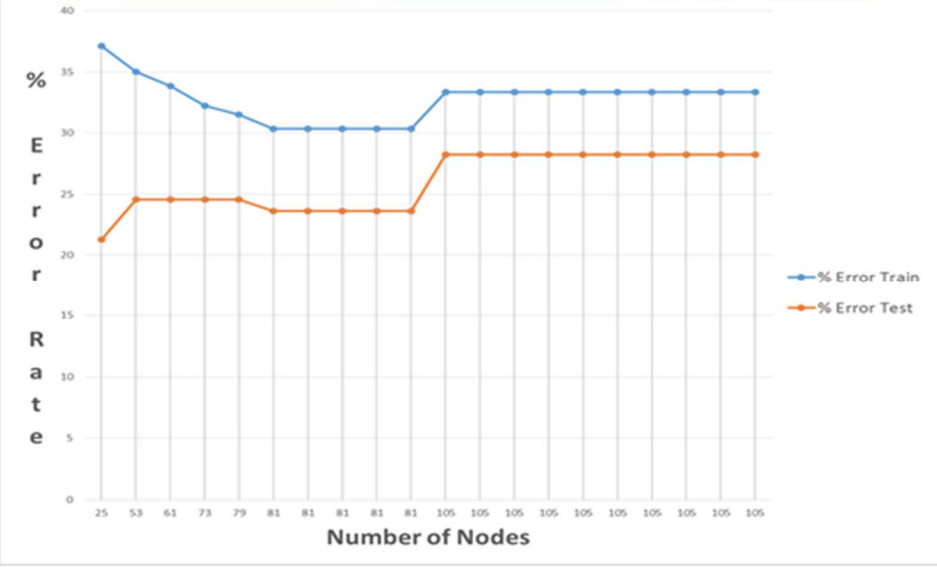
Figure 7 Obtained Values

Confidence Factor	Size of Tree	% Error Train	% Error Test
0.05	25	15.814	21.2963
0.1	53	10.4651	24.537
0.15	61	9.3023	24.537
0.2	73	7.6744	24.537
0.25	79	6.9767	24.537
0.3	81	6.7442	23.6111
0.35	81	6.7442	23.6111
0.4	81	6.7442	23.6111
0.45	81	6.7442	23.6111
0.5	81	6.7442	23.6111
0.55	105	5.1163	28.2407
0.6	105	5.1163	28.2407
0.65	105	5.1163	28.2407
0.7	105	5.1163	28.2407
0.75	105	5.1163	28.2407
0.8	105	5.1163	28.2407
0.85	105	5.1163	28.2407
0.9	105	5.1163	28.2407
0.95	105	5.1163	28.2407
1	105	5.1163	28.2407

Source: Author

The best confidence factor available for this dataset is 0.55 that indicates the best relationship among the tree size and error rate. Error rates closest at 105 nodes where confidence Factor is 0.55. This confidence factor was selected to use.

Figure 8. Erro rate x Number of Nodes



Source: Author

In conclusion, after the Classification Mining phase, the results showed that the best attributes are the attributes listed bellowed because they present high ratings.

Figure 9 Ranked Atributes

1.	RightToSmoke	0.992727273
2.	CannotAfford	0.980392157
3.	PhysiciansOnly	0.940758294
4.	SafetyPrecautions	0.932523617
5.	SafetyHazard	0.909090909
6.	EducationLevel	0.909090909
7.	NotCapable	0.893280632
8.	SecondHandHarmful	0.869722557
9.	StaffDuty	0.865451997
10.	CounselingFrustrating	0.8
11.	DesiredBehavior	0.75
12.	HarderToManage	0.674267101
13.	NotInterested	0.666666667
14.	NoneOfMyBusiness	0.666666667
15.	SmokingStatus	0.666666667
16.	MoreCare	0.628930818
17.	SafetyResidents	0.508905852

Source: Author

3 ASSOCIATION MINING

Associative analysis is an extremely useful method to find strong relationships in data sets. Through the Association it is possible to find useful links that are often not readily visible. These relationships can be represented by association rules that show items that are more frequent in this database. In consonance with Yin, Kaku, Tang and Zhu (2011), association rules mining in inventory database can help in many business decision-making processes such as catalog design, cross-marketing, cross-selling and inventory control.

In the Association phase, there were about 82 rules associated to the goal attribute among 1564 association rules. These 82 rules were filtered by some aspects, such as, if these ones have the main attribute in the second part of the if-rule. According to the Association phase definition, an interesting lift was settled in a value that is considered better than guessing. This special value is 0.99. A bunch of rules were generated and interesting conclusions were made.

Figure 9. Rules Generated

RightToSmoke=Agree	189 ==>	Position=Doctor_6	53 conf:(0.28)	<	lift:(1.09)>	lev:(0.01)	[4]	conv:(1.02)
SafetyHazard=Agree	194 ==>	Position=Doctor_6	55 conf:(0.28)	<	lift:(1.1)>	lev:(0.01)	[5]	conv:(1.03)
NoneOfMyBusiness=Neutral	135 ==>	Position=Doctor_8	51 conf:(0.38)	<	lift:(1.29)>	lev:(0.02)	[11]	conv:(1.12)
RightToSmoke=Strongly_Agree	142 ==>	Position=Doctor_8	53 conf:(0.37)	<	lift:(1.27)>	lev:(0.02)	[11]	conv:(1.12)

Source: Author

That was a bunch of important rules generated from the data set, all of them has the lift higher than 1. That means these rules are much better than guessing. These rules above can be translated as the four rules below:

1. If agree with they have the right to smoke the position is doctor 6
2. If agree with smoking is a safety hazard then position is doctor 6
3. If don't advise because is none of my business then position is doctor 8
4. If strongly agree with the right to smoke, then position is doctor 8

It's possible to say that Doctor 6 is someone who has some contradictory thoughts. In the first rule he believes that smoking is a right, but, in another hand, in the second rule, he believes that smoking is a safety hazard for the retirees. Maybe this position has an undefined opinion. About the doctor 8, he seems to be a very positive opinion about smoking. His answers generally say that he agree with the right to smoke but he does not have concern about this because in the third rule he says that does not advise the retirees because it is none of his business.

Other bunch of interesting rules is shown below:

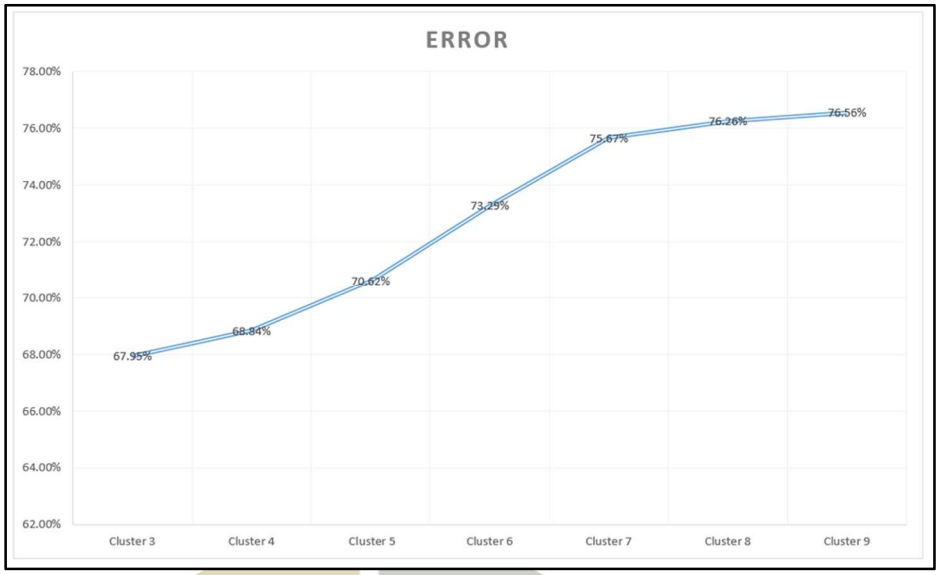
- If smoking status is never, then position is doctor 6
- If education level is 3 then position is doctor 8
- If education level is 4 then position is doctor 7
- If education level is 5 then position is doctor 6

Those rules generally say that the lower the number of the doctor the higher is their education level. That abroads some interesting conclusions if we compare these conclusion with the other found conclusions. We can infer that the higher the education level, the most concerned the professional is.

4 CLUSTER ANALYSIS

Clustering is defined as a phase of classification of patterns that can be called in different ways, such as observations, data items or features vectors. This phase has a main objective of organizing into groups, called clusters. Many disciplines have been come across clustering problems in distinct contexts, this emphasizes the great usefulness of clustering in data mining. (Jain, Murty, and Flynn, 1999) In this phase, the data is divided into clusters (groups). The elements of these groups are organized in similar criteria. On other words, similar elements are in the same group and distinct objects are in different groups. The objective in this phase, it is to extract information from a large amount of data. Specifically, in this project, the value of the main attribute Position were found as a result of some other attributes. There were computed 505 records from the previous phase and 16 attributes according to some criterias. These records were divided into a training set and a testing set, they contain 337 and 168 respectively. *SimpleKMeans* was the method chosen for grouping all of these records. This method as its name says, it is a simple way to organize all the records of the data in K distinct clusters. In addition, in this method the Euclidean Distance was applied in the data set. The results that were found showed some interesting features. As it was shown previously, there are 9 different values for the attribute Position in the data set. Some tests were made and the data was organized in rage between 3 and 9. Therefore, it was counted the error rate for each one of the clusters. The following graphic shows the error percentage.

Figure 10. Error Graph



Source: Author

It is clear in the graph that Cluster number 3 is the best one to analyze the data because this one does have the lower chance of wrong data, approximately 67.55%.

The following rules were extracted from analyzing the Cluster 3:

Figure 11. Rules from Cluster 3

Cluster 0
If SecondHandHarmful=Strongly_Agree
and rightToSmoke=Strongly_Agree
and StaffDuty=Strongly_Disagree
and NotCapable= Strongly_Disagree
and SafetyPrecautions=Strongly_Agree
and MoreCare=Strongly_Agree
and DesiredBehavior=Strongly_Disagree
and SafetyHazard=Strongly_Agree
and PhysiciansOnly=Strongly_Disagree
and NotInterested = Agree
and HarderToManage= Strongly_Disagree
and CounselingFrustrating=Strongly_Disagree
and NoneOfMyBusiness=Strongly_Disagree
and CannotAfford=Strongly_Disagree
and SafetyResidents= Extremely_Concerned
and EducationLevel= Level_3
and SmokingStatus= Never
Then position = Doctor 6

Cluster 2
If SecondHandHarmful= Agree
and rightToSmoke= Agree
and StaffDuty=Neutral
and NotCapable= Disagree
and SafetyPrecautions=Agree
and MoreCare=Neutral
and DesiredBehavior= Neutral
and SafetyHazard= Agree
and PhysiciansOnly= Neutral
and NotInterested = Neutral
and HarderToManage= Neutral
and CounselingFrustrating= Neutral
and NoneOfMyBusiness= Neutral
and CannotAfford= Neutral
and SafetyResidents= Extremely_Concerned
and EducationLevel= Level_3
and SmokingStatus= Never
Then position = Doctor 8

Source: Author

Above there are two of the three clusters, they shown interesting conclusion that match with the conclusions found on the Association phase. The doctor six still has the same contradiction found on the association phase. He agreed with the right to smoke but in another hand, he agreed that it is a safety hazard. In the second cluster, the doctor 8 agreed with the right to smoke and has a neutral opinion when he says that advising the retiree is none of his business.

CONCLUSION

In conclusion, all the work that has been done here showed interesting results. The rules generated in the end summarized important aspect about the opinion of the health staff about retirement home residents on tobacco cessation. All the phases have showed a variety of rules in the project. The preprocessing was important because there is a large amount of not useful data. This phase helped to clean the data for the next stages. After that, the classification phase helped to gather the important attributes that was better connected with our main attribute. This phase was essential because this important attributes were used as entries to the next two phases, these two phases were where the rules could be generated. Finally, those rules could be transformed in a simple language and final conclusions could be made to help the identification of professions who are more willing to advice old people who live in retirement home.

REFERENCES

GARCIA, Salvador, LUENGO, Julián, HERRERA, Francisco. **Data preprocessing in data mining**. New York: Springer, 2014.

GIBILISCO, Stan. **Statistics demystified**. 2.ed. New York: McGraw-Hill, 2011.

KUMAR, Senthil. **Knowledge discovery practices and emerging applications of data mining**: trends and new domains. Los Angeles: IGI Global, 2010

WITTEN, Ian H. Frank, EIBE. Hall, **Data mining**: practical machine learning tools and techniques Atlanta: Elsevier, 2011.

YIN, Yong. KAKU, Ikou. TANG, Jiafu. ZHU, JianMing. **Data mining:** concepts, methods and applications in management and engineering design**.** London: Springer, 2011.

João Emmanuel D'Alkmin Neves

Ex-bolsista do Programa Ciência sem Fronteiras.Tecnólogo em Análise e Desenvolvimento de Sistemas pela FATEC/Americana com Graduação Sanduíche em Computer Science pela SUNY - State University of New York. Possui Graduação em Design Gráfico pela UNIP. Atualmente é analista de sistemas na Empresa PersonalSoft e atuou como analista desenvolvedor na IBM Brasil através de Systemplan Sistemas Projetos Com. Ltda. Experiência em programação orientada a objetos, mobile multiplataforma, big data, data mining, engenharia de software e business intelligence. Desenvolve atividades de pesquisa sobre sistemas embarcados, arduino, internet das coisas, computação nas nuvens, inteligência artificial, arquivística e preservação digital.

Contato: jeneves@gmail.com

Fonte: CNPQ – Curriculo Lates

Antônio André Cunha da Silva

Atualmente é pesquisador da Universidade Federal do Pará. Tem experiência na área de Ferramenta de Execução de Processos de Software Livre

Fonte: CNPQ – Curriculo Lates