Application of Knowledge Discovery Methodology in Public Data for Predicting Expenditures of Pharmaceutical Products Imports

Buzeti L[1,3] and Bortolozzi F[2,3] and Bernunci MP[2,4] and and Almeida IC[2,3*]

[1]Analista de Sistemas
Universidade Estadual de Londrina, Londrina, Brazil
buzeti@uel.br

[2]Bolsista do Programa Produtividade em Pesquisa do ICETI
Instituto Cesumar de Ciência, Tecnologia e Inovação

[3]Programa de Pós-Graduação em Gestão do Conhecimento
Centro Universitário de Maringá, Paraná, Brazil
flavio.bortolozzi | iara.almeida@unicesumar.edu.br

[4]Programa de Pós-Graduação em Promoção da Saúde
Centro Universitário de Maringá, Paraná, Brazil
marcelo.bernuci@unicesumar.edu.br

ABSTRACT

In the period from 2005 to 2014, Brazil had a trade deficit of US $ 16.074 billion related to drug trade. This financial effort to promote access to medicines by the Brazilian population had the purpose of avoiding morbidities and, consequently, avoids hospitalizations. Therefore the present study aimed to contribute to improve knowledge management in health, applying a new knowledge management model (Buzeti´s model), which allows the discovery of knowledge in public data of the ALICEWeb system. Initially, the existing methodologies that enabled the discovery of knowledge were studied, and the Buzeti`s model was defined. It should be noted that concepts derived from Knowledge Discovery in Database (KDD) have been chosen, which allow the extraction of patterns, associations, rules, clusters and other forms of coding. The steps proposed by the Buzeti`s model were performed: selection of ALICEWeb system public data; data pre-processing and cleaning; transformation of data; data mining; and analysis and interpretation of ALICEWEB system information. The application of Buzeti´s model guaranteed the accomplishment of the drug import forecast for the year 2016. We conclude that the use of the KDD methodology allows the creation of new knowledge from public data and the Buzeti´s model can be used to improve the knowledge management in health.

Keywords: data mining, knowledge discovery in databases, public data

1 INTRODUCTION

Throughout the world economic changes, the traditional factors of production - labor, capital and land - have acquired secondary importance. In contrast, the knowledge factor is increasingly becoming the main resource of organizations [4]. According to Paton et al. [11], information and knowledge are the

most competitive weapons for companies, and the accumulation of quality knowledge provides a competitive edge within the market. Knowledge Management (KM) aims to assist organizations in dealing with this asset. According to Dalkir [2], KM is:

> "the deliberate and systematic coordination of people, technology, processes and organizational structure to add value through reuse and innovation. This coordination is achieved through the creation, sharing and application of knowledge as well as by feeding the valuable lessons learned and best practices to corporate memory in order to promote continued organizational learning"

With such importance to the business world, the need to manage knowledge is obvious and demands complexity. According to Davila et al.[3], the complexity of KM can be observed from the cycles that form it, as the capture (creating and retrieving), sharing (disseminating) and applying (using) knowledge in the organization. Figure 1 shows how these cycles occur. It should be emphasized that if these cycles are duly deployed and matured within an organization, business excellence can be fostered because the registered and shared knowledge allows for greater efficiency and productivity, prevention of mistakes, and ensures better performance of the innovation process.

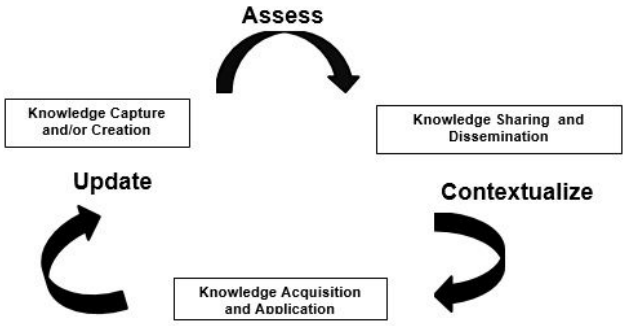

Figure 1. Cycle of Knowledge Management (adapted from [2]).

This search for innovation and the consequent incorporation of the knowledge in the generated products, classifies the industrial sectors in levels of technological intensity. One way to identify such levels is by analyzing average spend on research and development on billing. This expenditure is one of the parameters that define the technological intensity of each industrial sector, since it increases the level of knowledge incorporated to the products [16].

The search for innovation and the consequent incorporation of knowledge into the products generated, classifies the industrial sectors into four levels of technological intensity: high, medium high, medium low and low. One of the ways to identify such segments is by analyzing the average spends on research and development on billing. This spending is one of the parameters that define the technological intensity of each industrial sector, since it increases the level of knowledge incorporated into the products [16].

The pharmaceutical industry is classified in this segment of high technological intensity precisely because it has the characteristics required. This innovative effort results in knowledge externalities, both tacit and explicit knowledge. These externalities, through the production of new knowledge, result in the country development and inflow of financial foreign exchange. However, in the Brazilian case, there is a continuous deficit in the pharmaceutical trade between the country and the world. In the data related to external transactions (import / export) of drugs, it is observed that, in 2005, Brazil imported 3.31 billion dollars. In the year 2011 the country imported 9.50 billion dollars [6]. However, this phenomenon cannot be observed only by the financial bias, but also by what produces of social well-being.

Despite this international recognition for the large amount of data, the ALICEWeb system can be considered in what Han and Kamber [6] describe as being "data-rich, knowledge-poor." Through the analysis of this data set, the knowledge management, using the cycles of capture and / or knowledge creation, can contribute to the improvement of Brazilian public health management. This creation and / or capture of new knowledge can assist health managers in the processes of innovation, planning and decision making. Using knowledge management tools and techniques, it is possible to systematically extract code and organize knowledge [2]. It is possible to use tools that can analyze the relationship between variables (drugs, hospitalizations and others) within a large volume database, in order to extract new knowledge. Thus, the challenge of providing quality health services to the Brazilian population may have knowledge management as an ally. All resources, mainly technology-based, can be used for the purpose of improving the public health system.

With the development of information and information technologies (ICT), the process of extracting knowledge from large volumes of data and its complexity can be very expensive, depending on the characteristics of the problem to be solved. In organizations, the amount of data currently generated surpasses the human capacity to interpret and understand such a large amount of information. In order to address this problem, within the Computing area, the research sub-area known as: Knowledge Discovery in Databases (KDD) [5] emerged. The KDD is an essentially interdisciplinary technology involving primarily the areas of: Databases, Artificial Intelligence (Neural Networks, Fuzzy Logic, etc.) and Statistics, among others.

Therefore, from a large amount of data and information, in the most diverse areas of knowledge, managers and health professionals involved in decision making demand for new possibilities to optimize their process. In this way, the following research question emerges: Knowledge Management Tools can collaborate to understand the relationship between the variables "drug importation" and "morbidity index" of the Brazilian population?

Thus, the present study aims to improve the management of information and knowledge in the health area. This contribution focuses on the extraction of new knowledge that can be used for the benefit of Brazilian public health. This knowledge will be created from the understanding of the relationship between health and import variables. The use of technological tools of knowledge management as the process of Knowledge Discovery in Databases will allow the extraction of such

knowledge. Therefore the general objective of this study is the creation of new knowledge, using KDD techniques for data extraction from the ALICEWeb system.

2 CONCEPTUAL PREMISES

In the course of the present study, an analysis was made of the main concepts, definitions and techniques used throughout the development of the Buzeti´s model: knowledge discovery practices in database, data mining, big data, knowledge management, understanding of the ALICEWeb site and statistical concepts.

Taking this scenario into account, Silva and Breternitz [15] define big data as "a set of technological trends that allows a new approach to the treatment and understanding of large data sets for decision-making purposes." From the need to analyze, in a non-traditional way, large amounts of data, the concept of Knowledge Discovery in Databases (KDD) arises. According to Rezende [14], the KDD consists of the phases:

1. **Data selection** - the data set is chosen with its variables, attributes and records. The source (s) for this selection may come in different formats, such as data warehouses, spreadsheets or information systems;
2. **Pre-processing and data cleaning** - the aim is to ensure data quality by eliminating redundant and / or inconsistent data, retrieving incomplete data and evaluating outliers;
3. **Transformation of data** - data is formatted and stored correctly;
4. **Data mining** - allows the discovery of new knowledge;
5. **Interpretation and evaluation** - the knowledge generated is evaluated by specialist (s).

According to Prass [13], the process can return to one of the previous phases if the evaluation phase considers that the knowledge generated is not adequate. According to Lemos [8], Data Mining uses concepts from both the statistical area and the Artificial Intelligence area, more specifically from the Machine Learning subarea. Artificial Intelligence, according to Barr [1], is the part of Computer Science focused on the development of intelligent systems. Machine learning, according to Michalski et al [10], intends that a program learns and improves its performance given the experience gained in a particular practice. It is concluded that Machine Learning differs from traditional statistical methods, mainly due to the concepts of confidence intervals and standard errors.

Pinheiro et al [12] affirm that Health Information Systems "still do not reach their full potential as they are used in an incipient way by health management for the decision-making process". It is known that effective knowledge management can be done through the use of information technology. In addition, large amounts of data and information available can become a great ally for managers, since they are converted into knowledge, an essential asset for the improvement of systems and organizations. Thus, the challenge of providing quality health services to the Brazilian population may have Knowledge Management as an all.

The ALICEWeb system of the Brazilian Ministry of Development, Industry and Foreign Trade was developed in order to modernize the access to the statistics of exports and imports. This system is updated monthly with data from the previous month and the data sources come from the Ministry of Development, Industry and Foreign Trade (MDIC). MDIC's mission is: "To formulate, implement and evaluate public policies to promote the competitiveness of foreign trade, investment and innovation in

business and consumer welfare" [9]. Thus, any commercial transaction between Brazil and other countries of the world is properly registered in the system ALICEWeb.

3 METODOLOGY

It is a study of the development of applied methodology with the use of secondary data sources. We performed procedures of bibliographic and experimental research, applying the case study in the ALICEWeb system. In order to achieve the general objective of this study, which is the creation of knowledge from a public database, the following research questions were elaborated:

1. What are the concepts of Knowledge Management and technological that will be necessary to base the research?
2. How to propose a model to solve the problem proposed?

In order to answer the first question, searches were performed on electronic databases, both Brazilian and international, using the words "knowledge management", "data mining", "knowledge management", and "data mining". The databases used were Google Academics, LILACS and MEDLINE. Only relevant studies published between 2009 and 2015 addressing public health data mining were selected. The articles were initially selected by titles and abstracts and then by the review of the full article. To answer the second question, a KDD Model based on Fayyad [5], called the Buzeti Model, was proposed. This model has five steps that correspond to the phases of knowledge discovery for ALICEWeb System data: (1) Data selection, (2) Pre-processing and cleaning of data, (3) Data transformation, (4) Data Mining, e (5) Analysis and Interpretation of ALICEWEB system information.

4 RESULTS

First, we present the stages of development of the Buzeti´s model:

Step 1 - Data selection: Specifically in this study, data from Chapter 28, entitled "Inorganic chemicals, inorganic or organic compounds of precious metals, radioactive elements, rare earth metals or isotopes" were explored.

Step 2 - Preprocessing and data cleaning: 42 different tables were generated using the "filter by Chapter 28" option. The features in this option is that the system allows a maximum of 6 periods per query. Thus, since the study had a period of coverage of 84 months, it was necessary to repeat the query 14 times, which resulted in 42 queries / tables. After receiving the 42 emails from the ALICEWeb system, it was necessary to save the 42 tables that were attached. These tables originally had names that make it difficult to identify which chapter and period the table corresponds to. Therefore, to facilitate the process of grouping all tables into a single database, the following table naming standard was adopted: "capXXYYZZAAAA" so that XX indicates the Chapter number, YY starting month, ZZ final month And YYYY indicates the year.

Step 3 - Data Transformation: In order to solve the difference between the semesters, we chose to check the intersection of the two tables, the first and the second semester, and maintain the drugs that were common to both. Drugs that were not common were eliminated and separated into a waste table. Twelve drugs were eliminated in "cap2801062009" and in table "cap2807122009" were eliminated 17. This operation was repeated for all pairs of tables, of the first and second semesters. It was possible to note that 29 (17 + 12) drugs that were not included in one of the tables were eliminated, thus, 29 drugs could not be analyzed, considering only the first two tables. It was also possible to note that from

the original consultation of 458 drugs in Chapter 28, 327 remained after the intersection of the two tables. This represented a difference of 131 drugs, which were also not analyzed. The combination of the semiannual tables generated annual tables. The annual tables of chapter 28 did not have the same amount of drugs, which made it impossible to join in order to form only one table containing the 7 years of coverage of the present study. As occurred in the semiannual tables, we chose to verify the intersection of the seven (annual) tables and, from the drugs that were common to both tables, to keep them in each of the tables. Those that were not in common were disposed of and separated into a waste table. In this way, the tables were concatenated, which generated a final database with 295 drugs.

Step 4 - Data Mining: In this step, several modeling techniques were selected and applied as well as the parameters of the forecast models. There are several techniques for the same type of data mining problem, techniques that were performed within the R software. The following techniques were performed: analysis of time series with trend line, and forecast for dollar expenditures through training and data testing.

We compared all the results obtained in order to find the best models that resulted in the creation of new knowledge and that contributed to health management.

Step 5 - Analysis and Interpretation of ALICEWeb system information: The results obtained by Data Mining were analyzed. After interpreting that the time series is not stationary, it does not have seasonality but has a tendency, the best model was used to carry out the forecasts, taking into account these characteristics. Although there are five types of trends in the literature, we identified in the time series the presence of an additive tendency, justified by the fact that the trend variability is discrete. More precisely, there was no marked and gradual increase or decrease in data over time, which evidenced the non-multiplicity but the additive tendency. With the methodology of separating the data into two parts, one for training and another for testing, seven models of prediction were tested, as shown in Figure 2. The dashed line represents the test data (or actual data). This line was compared to the other lines representing the forecasting methods. It was identified that the Naïve, Mean and Holt Alpha methods are linear and contrary to the test data, being in a position far from the test data. This indicated a poor performance compared to forecasts.

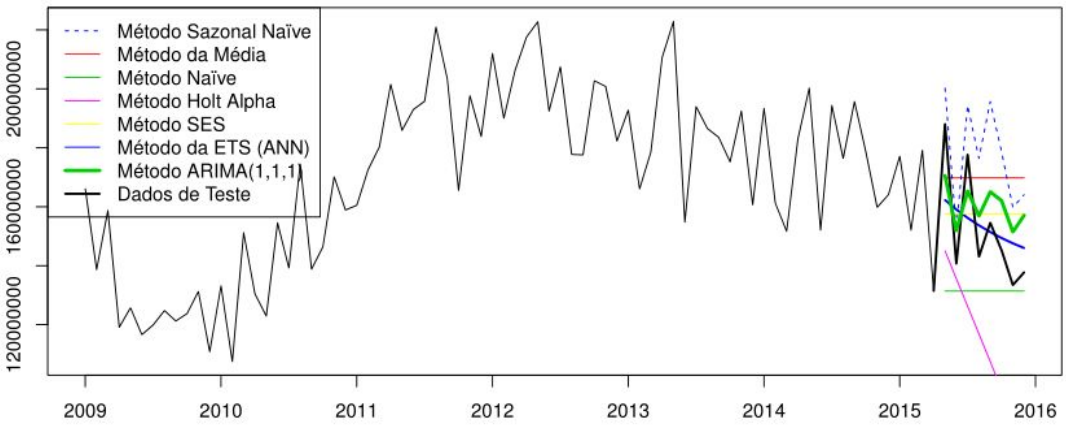

Figure 2 - Testing forecast models. Source: The Authors

To evaluate the other methods we tried to elaborate the prediction errors for each of the methods. This evaluation is presented in Table 1. Thus, by analyzing both Figure 5 and Table 1, we opted for the ARIMA method (111) (001) because it has the lowest MAPE and the lowest MASE and, consequently, the lowest prediction error in relation to the test data.

	ME	RMSE	MAE	MPE	MAPE	MASE	Type
7	7881795.86	30623184.78	24654116.05	3.57	14.15	1.00	Snaive
3	0.00	29843883.13	25084390.66	-3.49	16.00	1.02	Média
1	-463443.20	24019663.57	19964486.80	-1.38	12.14	0.81	Naive
6	1488962.70	21297242.69	17382690.47	0.39	10.56	0.71	HoltAlpha
5	721458.32	18971518.35	15632813.78	-0.61	9.43	0.63	SES
4	230843.18	18203326.08	14502063.43	-0.56	8.77	0.59	ETS(A,N,N)
2	-205145.50	17327973.94	14121535.85	-0.92	8.53	0.57	ARIMA(111)(001)

Table 1 - Prediction errors for each of the forecast methods. Source: The Authors

The data referring to the prediction of expenditures in dollars with drugs using the Buzeti´s model are presented in figure 6. It should be noted that the predictions, represented in the blue line, with 80% review intervals in the dark gray shaded area and the 95% prediction intervals in the light gray shaded area.

5 CONCLUSION

This study presents the development of the Buzeti´s model directed to the preparation and analysis of public data regarding the importation of drugs from the Brazilian AliceWeb system. This model uses techniques presented by KDD and consists of five steps: data selection, data pre-processing and cleaning, data transformation, data mining and analysis and interpretation of ALICEWeb system information.

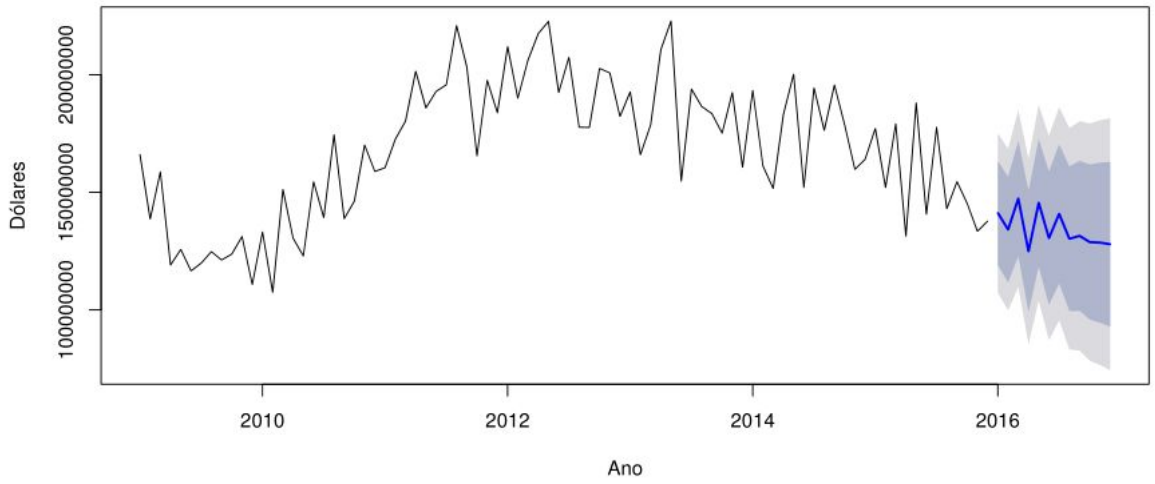

Figure 06 - Expenditure forecast with drug imports applying the Buzeti model. Source: The Authors

The Buzeti model explains the complexity inherent in the public database of the ALICEWeb system and presents a procedure for the construction of a single database, from the joining of dozens of tables generated by the ALICEWeb system. This study highlights the differentiated amount of drugs per year, as well as the difficulty in working with the data because of its complex and sometimes incomplete structure. For the forecasting of expenses with the importation of drugs for the year 2016, 7 forecasting methods were studied - to evaluate the data in the period of 2012 to 2015 - the ARIMA method (111) (001) was chosen. Therefore, our results contribute to future studies that need to follow a methodology for acquiring new knowledge from public data of the ALICEWeb system.

6 REFERENCES

[1] **BARR, A. Feigenbaum**. 1981. "*the handbook of artificial intelligence*,"volumes i and ii, william kaufmann. Inc., Los Altos, Ca.

[2] **DALKIR, K.** 2005. Knowledge *Management in Theory and Practice*. Elsevier Science. ISBN 9780080547367. Disponível em: <https://books.google.com.au/books?id=xtFLTymKV0QC>.

[3] **DAVILA, G. A. et al.** 2014. *O ciclo de gestão do conhecimento na prática: Um estudo nos núcleos empresariais catarinenses*. International Journal of Knowledge Engineering and Management (IJKEM), v. 3, n. 7, p. 43–64.

[4] **DRUCKER, P. F.; DRUCKER, P. F.** 1994. Post-capitalist society. [S.l.]: Routledge.

[5] **FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P.** 1996. *The kdd process for extracting useful knowledge from volumes of data*. Communications of the ACM, ACM, v. 39, n. 11, p. 27–34.

[6] **HAN, J.; PEI, J.; KAMBER, M.** 2011. Data mining: concepts and techniques. [S.l.]: Elsevier.

[7] **INTERFARMA.** 2012. *Balanço das políticas industriais para o setor farmacêutico.* Interdoc VOLUME III.

[8] **LEMOS, E. P.; STEINER, M. T. A.; NIEVOLA, J. C.** 2005. *Análise de crédito bancário por meio de redes neurais e árvores de decisão: uma aplicação simples de data mining*. Revista de Administração, v. 40, n. 3, p. 225–234.

[9] **MDIC.** 2017. *Ministério da indústria, comércio exterior e serviços: Institucional*. Disponível em: <http://www.mdic.gov.br/institucional>.

[10] **MICHALSKI, R. S.; CARBONELL, J. G.; MITCHELL, T. M.** 2013. *Machine learning: An artificial intelligence approach*. [S.l.]: Springer Science & Business Media.

[11] **PATON, C. et al**. 2015. *O uso do"balanced scorecard"como um sistema de gestão estratégica*. Revista de Ciências Jurídicas e Empresariais, v. 1, n. 1.

[12] **PINHEIRO, A. L. S. et al.** 2016.*Health management: The use of information systems and knowledge sharing for the decision making process*. Texto & Contexto-Enfermagem, SciELO Brasil, v. 25, n. 3.

[13] **PRASS, F. S.** 2007. *Kdd-uma visal geral do processo*.

[14] **REZENDE, S. O.** 2003. *Sistemas inteligentes: fundamentos e aplicações*. [S.l.]: Editora Manole Ltda.

[15] **SILVA, L. A.; BRETERNITZ, V. J.** 2013. *Big data : um novo conceito gerando oportunidades e desafio*. Revista Eletrônica de Tecnologia e Cultura, Jundiaí, v. 13, p. 106-113.

[16] **ZAWISLAK, P. A.; FRACASSO, E. M.; TELLO-GAMARRA, J.** 2013. *Intensidade tecnológica e capacidade de inovação de firmas industriais*. Proceedings of the ALTEC.