BRAIN. Broad Research in Artificial Intelligence and Neuroscience
ISSN: 2068-0473 | e-ISSN: 2067-3957
Covered in: Web of Science (WOS); PubMed.gov; IndexCopernicus; The Linguist List; Google Academic; Ulrichs; getCITED; Genamics JournalSeek; J-Gate; SHERPA/RoMEO; Dayang Journal System; Public Knowledge Project; BIUM; NewJour; ArticleReach Direct; Link+; CSB; CiteSeerX; Socolar; KVK; WorldCat; CrossRef; Ideas RePeC; Econpapers; Socionet.

2020, Volume 11, Issue 2, Sup.1, pages: 99-106 | https://doi.org/10.18662/brain/11.2Sup1/97

Adapting CRISP-DM for Social Sciences

Mihaela CAZACU[1],
Emilia TITAN[2]

[1] The Bucharest University of Economic Studies, Romania, czc.mihaela@gmail.com

[2] The Bucharest University of Economic Studies, Romania, emilia_titan@yahoo.com

Abstract: The growth of available data in the social sciences led to numerous knowledge discovery projects being launched over the years. Even if the volume and the speed of data are increasing, in social sciences data has an important limitation in terms of methodological process that drives the conceptual and analytical questions posed to the data. Social sciences domain experiences several challenges in their desire of extracting useful and implicit knowledge due to its inherent complexity and unique characteristics, as well as the lack of standards for data mining projects. The aim of this research is to bring Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology as a standardization in analyzing large volumes of unstructured data to generate analytical insights for wellbeing and social sciences topics in general. Also, taking into consideration that for a data scientist, the most time-consuming activity is data preparation step, we are trying to make more efficient this process using a clear methodology and tasks. Conclusion is that using a strong methodology with well-defined steps in research can increase productivity in terms of time and enhance the quality of the research.

Keywords: *CRISP-DM; predictive modelling; quality of life; social sciences; analytics.*

How to cite: Cazacu, M., & Titan, E. (2020). Peculiarities of Providing Psychological Assistance to Abused Children. *BRAIN. Broad Research in Artificial Intelligence and Neuroscience, 11*(2Sup1), 99-106. https://doi.org/10.18662/brain/11.2Sup1/97

Introduction

Big data is mainly related to two general concepts: data mining (DM) and knowledge discovery in databases (KDD), concepts that are interchangeably within projects and the specialists that are using them. Sagiroglu and Sinanc's (2013) defined big data as "*massive data sets having large, more varied and complex structure with difficulties of storing, analysing and visualizing for further processes or results*". Main difficulty comes from velocity, data complexity and variability for data storage, analysis and visualization (Azevedo and Santos, 2008).

The borders of using big data sets for social sciences research were opened by Electronic SNs. They used some advanced data management systems as Apache Hadoop or other technologies, in order to help social scientists to access very easy big sets of data and generate useful insights, but the main issue in this approach remained the methodological process. Another major challenge was to adapt the methods to analyze the data, SNs providing big amount of data, since such research has traditionally relied on small samples of data using established methodologies (Dubitzky et al. 2007).

Problem statement

Data mining processes such as the Cross Industry Standard Process for Data Mining (CRISP-DM) have been developed as a guideline for data mining projects (Shearer, 2000; Wirth, 2000). However, such processes, developed prior to the 'data boom' age are without due cognizance to the amount and multi-structured nature of data generated by modern information systems (Karisik, E. (2018). The golden value of data is created when insights are derived from it to enable meaningful knowledge generation (Bosnjak, Grljevic and Bosnjak (2009)).

In 1989, appeared the term knowledge discovery in databases, referring to a complex process of finding meaningful information and knowledge in data, and to emphasize the "high-level" application of specific data mining methods (Fayyad et al, 1996). From Fayyad's perspective, data mining is just one phase of knowledge discovery process, basically data patterns discovery.

The knowledge discovery process (Fayyad et al, 1996) is the process of using data mining methods to extract what is knowledge according to the specification of measures and thresholds by using a database with any required pre-processing, sub-sampling, and transformation of the database using five major steps: data selection, pre-processing, transformation, data mining and evaluation.

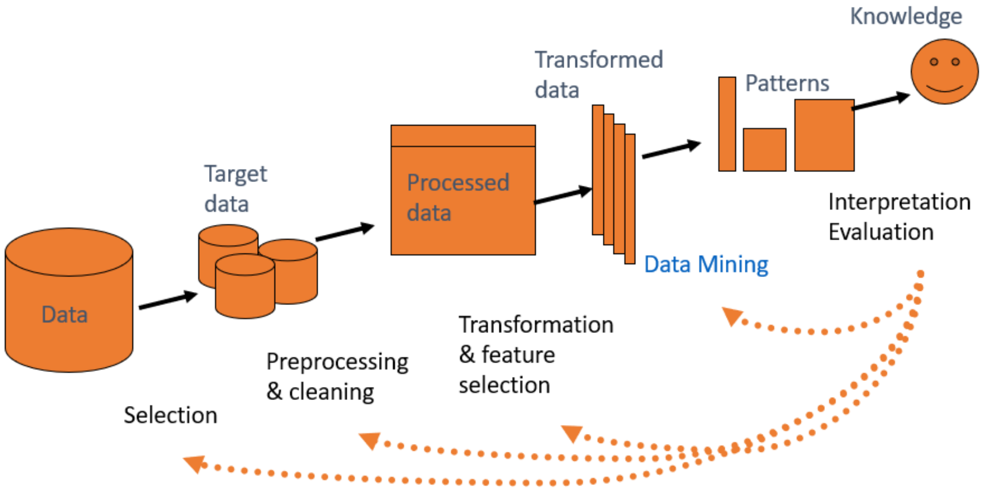

Figure 1. Fayyad's KDD Methodology
Source: Fayyad et al., 1996

Another methodology for knowledge discovery data, SEMMA, was developed on a later stage by the SAS Institute and it's an acronym which stands for Sample, Explore, Modify, Model, and Assess. As well as Fayyad methodology, SEMMA contains a list with sequential steps, guiding in this way the data mining projects development.

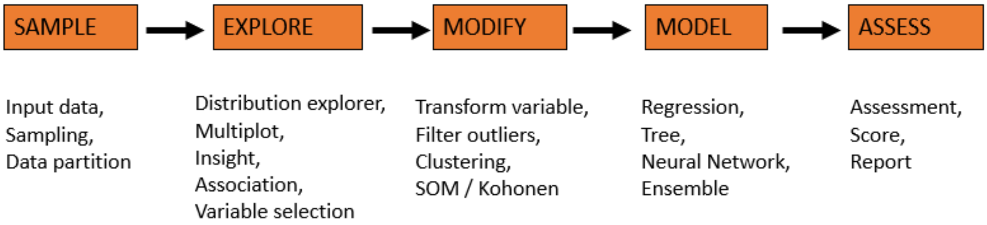
SAMPLE
EXPLORE
MODIFY
MODEL
ASSESS
Input data, Sampling, Data partition
Distribution explorer, Multiplot, Insight, Association, Variable selection
Transform variable, Filter outliers, Clustering, SOM / Kohonen
Regression, Tree, Neural Network, Ensemble
Assessment, Score, Report

Figure 2. SEMMA methodology

By using SEMMA methodology steps to be followed are: data sampling as an optional step, explore, model and modify the data, and the last one, asses the data.

In addition to the existing two methodologies, CRISP-DM is bringing something new: an overview of the life cycle of a data mining project that contains the corresponding phases of a project, their tasks, and relationships between these tasks (Chapman, et al., 2000).

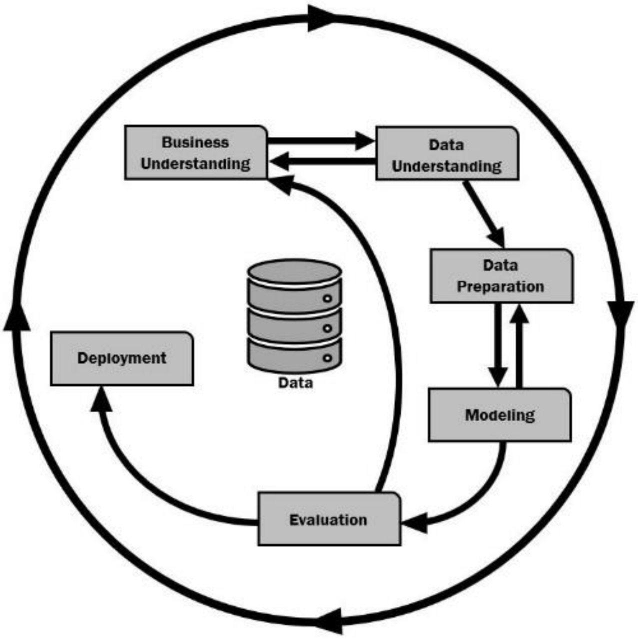

Figure 3. CRISP-DM methodology
Source: Chapman et al., 2000

As previous methodologies, also CRISP-DM comes with some general steps and more, tasks for each step: business understanding, data understanding, data preparation, data modelling, results evaluation and deployment (Vleugel, Spruit and Daal, 2010).

Now, the question is: how can CRISP-DM can help in social sciences from methodologically point of view? Can this methodology bring value in data knowledge discovery process and if yes, how?

Aims of the research

The main contribution of the CRISP-DM methodology adapted to social sciences, consists in defining clear steps and tasks specialized and to validate this we have built a data mining system and the entire system was applied to happiness related indicators provided by World Happiness report that supports the UN high level meeting on happiness and human wellbeing (Moreira, De Carvalho, Horvath, 2018). The happiness scores and rankings use data from the Gallup World Poll, where happiness is defined based on six factors - economic production, social support, life expectancy, freedom, absence of corruption, and generosity.

Crisp-DM methodology was adapted to this research exercise and enhanced with specific steps and tasks:

Step 1. Happiness, part of social sciences area, domain understanding

This step is referring to a very well understanding of happiness and wellbeing area (OECD, 2013). Mainly in the specialized literature are described two ways to evaluate the quality of life: subjective and objective (Frey and Luechinger ,2007). Between the two approaches are a lot of discussions which one is better and bring more insights in social research. Quantitative methods are referring to objective approach and had been developed a lot. In present this type of researches are dominating social sciences research landscape.

The major tasks for this step are:

• The researcher should understand the domain to ask the right questions

• Very well-understanding of World Happiness report, how the data is collected, spatial and temporal dimension, major social phenomena that might influence the results of the survey

Step 2. Data Understanding

To understand the data, we must do a simple analysis of the data sets like descriptive statistics for each feature included and to compute simple statistics like central tendency indicators, missing values in order to detect at first sight problematic features and anomalies. Specific tasks might include:

• Feature type: nominal, numeric, etc.

• Score meaning: respondents need to rate their life from zero to ten. Ten is the best life possible and zero is the worst

• Basic statistics for anomaly detection

• Conceptual correlations between features

Step 3. Dataset Preparation

At this step it is very important to create trust for data. In this direction, we are looking for completeness, consistency and accuracy in our data.

In our dataset we have no missing values and the values that we have are consistent. So, basic steps are:

- Data cleaning using different methods: custom substitution value, in the case of numeric features, replacing missing values with mean, remove entire column if needed
- Defining threshold filters like less than, in range, out of range, etc.

Step 4. Data Modelling

At this step, we need to look again at our dataset and to decide if it is the case to reduce the number of feature and delete some columns or to define new columns by merging others (Sarkar, Dipanjan & Bali, Raghav & Sharma, Tushar, 2018). Also, we need to define which will be our target variable for our happiness rank, and it will be the score, because score contains the cumulative information for the dimensions and rank it's just the consequence of the score.

Specific tasks might be:

- Data manipulation: create new columns based on computation of two or more columns already present in dataset
- Filter based feature creation based on predictive power
- Split the rows into two datasets: test and train data and the proportion is following Pareto principle (20 – 80) or 30-70
- Train the regression model based on our target column: happiness score
- Score the model: scores predictions

Step 5. Model evaluation and results interpretation

Model evaluation is the essential part of any analysis even if we are talking about social sciences or other domain. Now there are generated some standard metrics based on model choose at previous step. Specific tasks are:

- Interpreting the metrics generated by the model
- If the results are not the expected ones, we have to go again to step 4 and refine the model used or features selected

Step 6. Overall model analysis

At this step is important to look again to the model and if something else can be added now it is the moment having the model big picture.

Conclusion

To conclude, in social sciences area especially, data started to be the gold for the complex analysis part. In order to fulfil the objective, we must ensure that data is correct and vey well structured, otherwise we fail.

In our experiment, the results were good, we have noticed how a step by step approach with properly defined tasks can reduce time consuming research and on the other hand the potential patterns found in data. For the researcher the process using a clear methodology and tasks enhance the work and time.

References

Azevedo, A., & Santos, M. F. (2008, January). *KDD, SEMMA and CRISP-DM: a parallel overview*. IADIS European Conference Data Mining, 182–185. http://recipp.ipp.pt/handle/10400.22/136

Bosnjak, Z., Grljevic, O., & Bosnjak. S. (2009). CRISP-DM as a framework for discovering knowledge in small and medium sized enterprises' data. In *2009 5th International Symposium on Applied Computational Intelligence and Informatics*, Timisoara (pp. 509-514). https://doi.org/10.1109/SACI.2009.5136302

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C. & Wirth, R. (2000*). CRISP-DM 1.0 Step-by-step data mining guide ().* The CRISP-DM consortium .

Dubitzky, Werner & Granzow, Martin & Berrar, Daniel. (2007). *Fundamentals of Data Mining in Genomics and Proteomics*. 10.1007/978-0-387-47509-7.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). *From Data Mining to Knowledge Discovery in Databases*. AI Magazine, 17(3), 37. https://doi.org/10.1609/aimag.v17i3.1230

Frey, B. & Luechinger, S. (2007). *Concepts of happiness and their measurement.* Metropolis Verlag.

Karisik, E. (2018). *A standardized Data Mining Method in* Healthcare – a pediatric intensive care unit case study. Utrecht University, Information and Computer Science.

Moreira, J., De Carvalho, A., & Horvath, T. (2018). *What Can We Do With Data?* In J. M. Moreira, A. C. P. L. F. De Carvalho, & T. Horváth (Eds.), A General Introduction to Data Analytics (Chapter 1). Wiley Online Library. https://doi.org/10.1002/9781119296294.ch1

Organisation for Economic Co-Operation and Development (OECD). (2013). *OECD guidelines on measuring subjective well-being*. Paris, France: OECD Publishing. 10.1787/9789264191655-en

Sagiroglu, S. and Sinanc, D. (2013) *Big Data: A Review*. 2013 International Conference on Collaboration Technologies and Systems (CTS), San Diego,20-24 May 2013, 42-47. https://doi.org/10.1109/CTS.2013.6567202

Sarkar, Dipanjan & Bali, Raghav & Sharma, Tushar. (2018). *Building, Tuning, and Deploying Models*. https://doi.org10.1007/978-1-4842-3207-1_5

Vleugel, A., Spruit, M., & Daal, A. (2010). *Historical Data Analysis through Data Mining From an Outsourcing Perspective: The Three-Phases Model*. International *Journal of Business Innovation and Research, 1*, 42-65. https://doi.org/10.4018/jbir.2010070104