JOURNAL OF ADVANCEMENT IN MEDICAL AND LIFE SCIENCES
Journal homepage: http://scienceq.org/Journals/JALS.php
ISSN: 2348-294X V7I4.01
Research Article Open Access

Construction of a Datamart and knowledge extraction using the decision tree application to medical data

Richard KANGIAMA,
Department of Basic Sciences, Faculty of Oil and Gas, University of Kinshasa, Democratic Republic of the Congo.
*Corresponding author: Richard KANGIAMA, E-mail: richard.kangiama@gmail.com
Received: April 04, 2020, Accepted: May 01, 2020, Published: May 01, 2020.

ABSTRACT:
Our contribution relates to the implementation of a DataMart, finally to carry out a study on the influence of the prenatal factors which contribute to the birth of the children, We took the case of the maternity of the Saint Joseph Hospital (during 2011, 2012) on the basis of prenatal consultation and other vital factors as well as the birth conditions of the children. We started by collecting the data in the different registers and building a data source in Excel. Then, we created our operational database in Access, and then build our mini data warehouse, based on the results of the data. We used extraction and data mining software which is SPAD to highlight the parameter which influences the weight of children at birth for example and other variables.
Keyword: *DataMart, Data mining, Datamart, Data warehouse.*

INTRODUCTION:
Decision-making is a key problem that preoccupies business managers. This decision-making involves modeling the various problems they encounter in management, hence the need for a model based on the decision tree. The mini data warehouse / datamart being a centralized and universal vision of all the information of the company. It is a structure which aims, unlike databases, to gather company data for analytical purposes and to help the manager in strategic decision-making. A strategic decision is an action taken by business decision makers who aim to improve, quantitatively or qualitatively, the performance of the company. A problem of knowledge extraction consists in extracting knowledge from a data warehouse or from another data source using the techniques of Datamining (decision tree, Bayesian networks, neural networks, etc.) by applying them through software such as SPAD, XSLT or other data analysis.

2. METHODOLOGY:
Construction of datamart / Mini data warehouse
We have modeled the operational database with the MERISE method which is considered as our data source which will feed our datamart. After preliminary analysis, we have released the following tables for the logical model of our operational database, the construction of the latter also goes through these different stages according to Raph Kimball.

Step 1: Define the process to analyze
The procedure or function refers to the subject of our mini data warehouse, We determine the business process of Saint Joseph Hospital concerned by our studies for our case, the process is the results of report on the deliveries of women at hospital.

Step 2: Determine the level of granularity of the data
Choosing the grain means deciding exactly what a record in a fact table represents. For example, the birthing entity represents the facts relating to each birthing and becomes the fact table of the star diagram of childbirths. Therefore, the grain of the childbirth fact table is a childbirth performed in the maternity ward. After choosing the grain of the fact table, we will begin to identify the dimensions of the fact table. By way of illustration, the entities sheet and obstetric history will serve as references to the data concerning deliveries and will become the tables of dimensions of the star diagram of deliveries. We also add Time as the main dimension; because it is always present in the star diagram based on whatever all the events take place in a very specific period.

Step 3: choose the dimensions
The dimensions determining the context in which we can ask questions about the facts established in the fact table. A well-constructed set of dimensions makes the mini data warehouse understandable and simplifies its use.
We identify the dimensions with sufficient detail, to describe things such as childbirth and properties with correct granularity.
For example, any person of the SHEET dimension is described by the attributes NF, NOMPOSTNOM, AGE, CIVIL STATUS, NATIONALITY, ADDRESS, LEVELS OF STUDY; the ANTOBSTETRI dimension is described by the following attributes: na, GRAVIDA, PARITY, ABORTION, DEATH; the TIME dimension is described by the following attributes TIME, DAYS, MONTHS AND YEAR.
Step 4: identify the metrics (facts): In our case, the fact is DELIVERY, the metrics are the digital data PROVENANCE (CPN), ETATSER, WEIGHT, AP GAR. Metrics are measurable and computable entities for statistical and numerical purposes.

3. RESULT:
Datamart: In summary, we can synthesize our decision system as such. Given that we are building a datamart we wish to stop at this stage, to explain, justified our case studies.

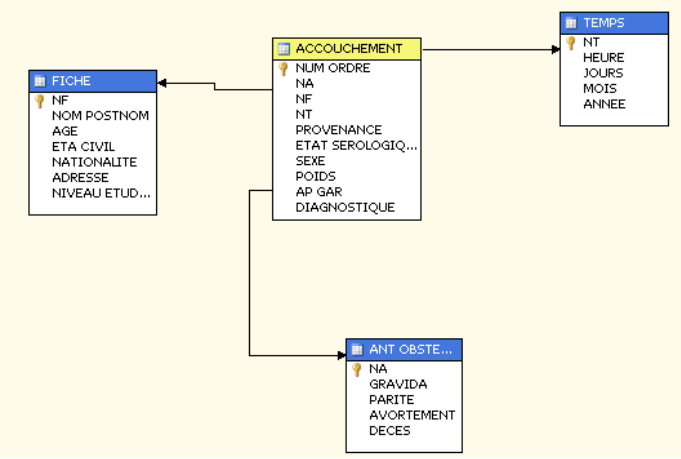

Figure 1: Star diagram of the Datamart.

5. Datamining module
To carry out our data mining module with the SPAD data analysis software, to facilitate our interpretation. We imported from an Excel file the result of an MDX query (SQL query on

the Multidimensional database) on our DataMart from there, we imported this data to SPAD to make the ACP, the complexity of these SPAD software and that it brings together a range of algorithms for analysis and data processing integrated within it. This table brings together the synthetic data obtained after our MDX query on the Datamart.

Editeur de données - memofin

Fichier Edition Affichage Fenêtre ?

	Iden	Libl	CPN	GEST	PAR	AVOR	POID
1	000001	Tr1	70.000	2.080	55.000	38.000	2886.300
2	000002	Tr2	124.000	2.250	108.000	50.000	2898.600
3	000003	Tr3	41.000	2.350	50.000	20.000	2919.700
4	000004	Tr4	135.000	2.280	250.000	48.000	2822.200
5	000005	Tr5	51.000	2.300	108.000	10.000	2809.100
6	000006	Tr6	224.000	2.280	529.000	82.000	2892.500
7	000007	Tr7	32.000	2.290	92.000	16.000	2759.400
8	000008	Tr8	79.000	2.210	196.000	22.000	2909.700
9	000009	Tr9	111.000	2.650	306.000	35.000	2909.700
10	000010	Tr10	61.000	2.380	174.000	6.000	3066.500

Figure 2. Synthetic table of Data representation

Description of age groups by interval:

Age range 1: from 19 to 21 years old, Age range 2: from 21 to 2 3 years old, Age range 3: from 23 to 25 years old, Age range 4: from 25 to 27 years old, Range Age 5: 27-29, Age 6: 29-31, Age 7: 31-33, Age 8: 33-35, Age 9: from 35 to 37 years old, Age range 10: from 37 to over.

We will present the results of our analyzes carried out with the SPAD software on the data of deliveries. Table of Clean Values and Contribution of Axes.

BINARY CORRESPONDENCE ANALYSIS:

OWN VALUES

OVERVIEW OF CALCULATION ACCURACY: TRACE BEFORE DIAGONALIZATION .. 0.0397

SUM OF OWN VALUES 0.0397

HISTOGRAM OF THE FIRST 4 OWN VALUES

NUMBER	VALUE	PERCENT.	PERCENT.		
			OWN		CUMULATIVE	
1	0.0360	90.79	90.79	**********
2	0.0035	8.73	99.52	********
3	0.0002	0.47	100.00	*
4	0.0000	0.00	100.00	*

CONTACT DETAILS, FREQUENCY CONTRIBUTIONS ON AXES 1 TO 4

ACTIVE FREQUENCIES

FREQUENCIES	CONTACT DETAILS	CONTRIBUTIONS	COSINUS SQUARES
IDEN - LIBELLE COURT P.REL DISTO	1 2 3 4 0	1 2 3 4 0	1 2 3 4 0
CPN - provenance CPN 2.90 0.26	-0.47 0.19 -0.06 0.00 0.00	18.0 31.1 48.0 0.0 0.0	0.85 0.14 0.01 0.00 0.00
GEST - gestation of the m 0.07 0.01	0.05 -0.03 0.01 -0.05 0.00	0.0 0.0 0.0 99.9 0.0	0.41 0.14 0.01 0.44 0.00
PAR - mother's parity 5.83 0.43	-0.64 -0.12 0.01 0.00 0.00	66.6 25.1 2.4 0.0 0.0	0.96 0.03 0.00 0.00 0.00
AVOR - the number of times before 1.02 0.36	-0.45 0.38 0.10 0.00 0.00	5.8 43.6 49.6 0.0 0.0	0.57 0.41 0.03 0.00 0.00
WEIGHT - the child's weight 90.17 0.00	0.06 0.00 0.00 0.00 0.00	0.00	9.6 0.2 0.0 0.1 0.0	1.00 0.00 0.00 0.00 0.00

CONTACT DETAILS, CONTRIBUTIONS AND COSINUS SQUARES OF INDIVIDUALS

AXES 1 TO 4

INDIVIDUALS	CONTACT DETAILS	CONTRIBUTIONS	COSINUS SQUARES
IDENTIFIER P.REL DISTO	1 2 3 4 0	1 2 3 4 0	1 2 3 4 0
Tr1 9.53 0.03	0.16 0.08 0.01 0.00 0.00	6.8 16.1 10.8 8.0 0.0	0.81 0.18 0.01 0.00 0.00
Tr2 9.94 0.02	0.05 0.12 -0.02 0.00 0.00	0.6 40.5 17.0 12.7 0.0	0.13 0.85 0.02 0.00 0.00
Tr3 9.47 0.04	0.21 0.01 0.01 0.00 0.00	11.4 0.3 6.5 5.3 0.0	0.99 0.00 0.00 0.00 0.00
Tr4 10.17 0.01	-0.12 0.03 -0.01 0.00 0.00	3.8 3.4 3.2 6.9 0.0	0.92 0.08 0.00 0.00 0.00
Tr5 9.31 0.02	0.13 -0.04 -0.01 0.00 0.00	4.6 4.2 8.7 6.7 0.0	0.91 0.08 0.01 0.00 0.00
Tr6 11.65 0.18	-0.43 0.01 0.01 0.00 0.00	59.4 0.4 4.1 1.7 0.0	1.00 0.00 0.00 0.00 0.00
Tr7 9.06 0.03	0.16 -0.04 0.02 0.00 0.00	6.6 3.5 27.0 5.5 0.0	0.93 0.05 0.02 0.00 0.00
Tr8 10.02 0.00	0.01 -0.04 0.00 0.00 0.00	0.0 5.1 0.8 8.2 0.0	0.07 0.92 0.01 0.00 0.00
Tr9 10.51 0.02	-0.13 -0.05 0.01 0.00 0.00	5.1 8.1 1.8 55.6 0.0	0.87 0.13 0.00 0.00 0.00
Tr10 10.34 0.01	0.07 -0.08 -0.02 0.00 0.00	1.6 18.4 20.0 3.5 0.0	0.45 0.51 0.03 0.00 0.00

C.2. Decision tree (dedogram)

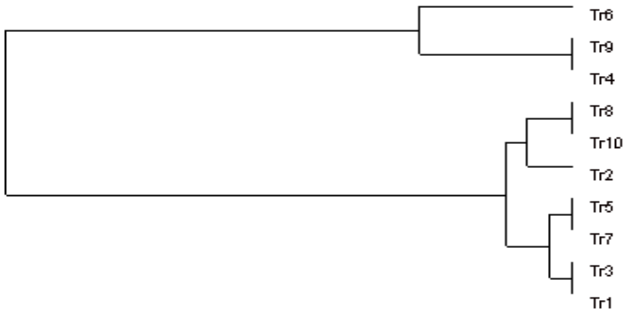

Figure 3. Dendrogram obtained after analysis with SPAD

C.3. Graphic

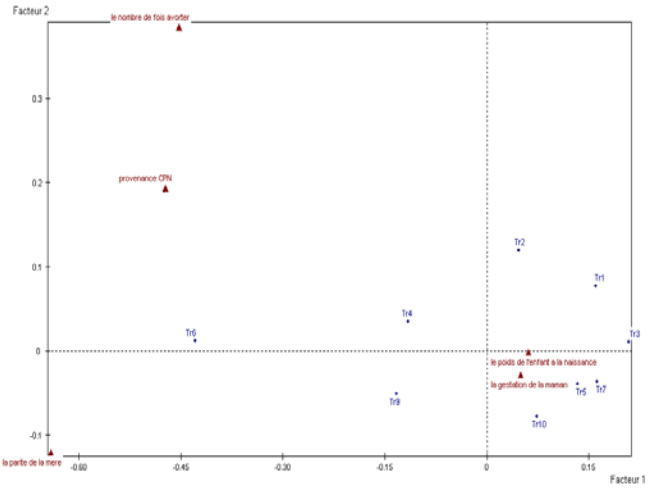

Figure 4. Variables grouping graph and contribution

6. Discussions on the results:

- Interpretation of results and discussion

Determination of axes

For the variables we take a 25% threshold:

We can say that for axis 1: The parity of the mother has contributed 66% to the creation of axis 1 it is of negative coordinates, for axis 2: The parity of the mother has contributed 25% when axis 2 was created, it has negative coordinates, which means that the number of times that the mother to give birth contributed 43.6% to the creation of axis 2, it has positive coordinates, prenatal consultation contributed 31.1% to the creation of axis 2, it has negative coordinates.

For axis 3: The prenatal consultation contributed 48% to the creation of axis 3 it has positive coordinates, the number of times that the mother gave birth contributed 49.6% to the creation of axis 3 is of positive coordinate.

For axis 4: Lagestite contributed 99% to the creation of axis 3 and has negative coordinates.
for individuals (10%)
For axis 1: Age group 3 contributed 11.4% to the creation of axis 1, it has positive coordinates and Age group 6 contributed 59.4% to the creation of axis1 it has negative coordinates.
For axis 2: Age group 1 contributed 16.1% to the creation of axis 2, it has positive coordinates and Age group 2 contributed 40.5% to the creation of axis 2 it has positive coordinates then, age group 10 contributed 18.4% to the creation of axis 2 it has negative coordinates
For axis 3: Age group 1 contributed 10.8% to the creation of axis 3 it has positive coordinates, age group 2 contributed 17% to the creation of axis 3 it has negative coordinates, age group 7 contributed 27% to the creation of axis 3 it has positive coordinates and age group 10 contributed 20% to the creation of axis 3 it has negative coordinates.
For axis 4: Age group 2 contributed 12.7% to the creation of axis 4, it has negative coordinates and age group 9 contributed 55.6% to the creation of axis 4 has positive coordinates.

Interpretation:

The age group going from 29 to 31 is associated with the parity of the mother or we can still say that the parity explains this age group better and the age group going from 29 to 31 is the age group whose women have given birth a lot, the age group 10 is associated with the ANC, so we can say that the majority of this woman has the ANC.
In this section 2, the age that mothers have aborted a lot.
We can still say in the age range from 21 to 23 years; young girls are often pressured into abortion.
In this section 7, the age when the majority of mothers have not followed the ANC. Most of these women believe they are already adults and neglect the ANC.
In this age group 9, this is the age group that the majority of women have already given birth to more than once.

4. CONCLUSION

Here we are at the end of our contribution, which focused on the extraction of knowledge from a DataMart using the decision tree, application to Medical data. In our work, we first spoke of the decision-making system which is presented as the set of processes which makes it possible to collect, integrate, model and present data. We also talked about the data warehouses which constitute the heart of the decision-making system playing a referential role for the company since it makes it possible to federate data often scattered in the different data sources. Finally with the notion of data mining allows for in-depth research on warehouse data with a decision tree for extracting knowledge. We performed the datamart with SQL Server 2008 with a decision tree model to allow us to make a decision on our data. From the above, we are convinced that all of the concerns respond to the problematic of our work. Our contribution was to carry out a DataMart on deliveries and from this; we were able to build a decision tree which we interpreted at the end.

REFERENCES

1. R. Kimbal, L. Reeves and W. Thornthwaite, the Data Warehouse: Guide to Project Management, Eyrolles, 4th edition 2008.
2. B. Burquier, Business Intelligence with SQL SERVER 2008: Implementation of a decision-making project, Dumond, 2008.
3. ADIBA .M, Data warehouses and data mining, Paris 2002;
4. Bertrand Burquier, Business intelligence with 2008, Implementation of a decision-making project, Dunod, 2009;
5. DANIEL T. LAROSE, From data to knowledge an introduction to Datamining, Vuibert, 2005;
6. GUIJARRO Vincent, The Decision Trees the ID3 algorithm, Lille, 2006;
7. KIMBALL .R and m. Ross, Data warehouses, practical guide to dimensional modeling, Vuibert, Paris, 2003;
8. N. KASORO, B. KASEREKA, Data warehouses and multidimensional analysis of the industrial exploitation of wood in the Democratic Republic of Congo, Annales faculté des sciences, Vol 1, year 2013.
9. Cristiano Xaviera, Fernando Moreiraa, *, Agile ETL, ERIS 2013 - Conference on ENTERprise Information Systems / PRojMAN 2013 - International Conference on Project MANagement / HCIST 2013 - International Conference on Health and Social Care Information Systems and Technologies, UniversidadePortucalense, Rua Dr Bernardino de Almeida, 541, 4200 Porto, Portugal, ScienceDirect
10. KasoroMulendaNathanael, KuyundaMayu Alain and NdoziMansangu R, excavating complex data and tuning the data warehouse a comparative study between the classic classification and the symbolic dynamic classification, the notebooks of the ISS-KIN, interdisciplinary journal, vol.9 , July 2014
11. Alain KUYUNSA MAYU *, Nathanael KASORO Mulenda **, RostinMatendo **, Modeling of Fuzzy datawerahouse, applied engineering, 2017
12. Calista M.Harbaugha, Jennifer N. Cooperb, Administrativedatabases, Volume 27, Issue 6, December 2018

Citation: Richard KANGIAMA (2020). Construction of a Datamart and knowledge extraction using the decision tree application to medical data, J. of Advancement in Medical and Life Sciences. V7I4.01. DOI: 10.5281/zenodo.3780416.