Evaluating Classification Algorithms performance with Matlab for generating alerts of risk of infant death

Gerson Albuquerque[1], Cristiano Silva[1], Joyce Quintino[1], Odorico Andrade[2], and Mauro Oliveira[1]

[1] Federal Institute of Education, Science and Technology of Ceará, Fortaleza, Ceará, Brasil
gersonvan@gmail.com
cristianocagece@gmail.com
joycequintino11@gmail.com
amauroboliveira@gmail.com

[2] Brazilian National Congress - Brasília, Distrito Federal, Brazil
odorico0811@gmail.com

Abstract

GISSA is an intelligent system for health decision making focused on childish maternal care. In this system, are generated alerts that involve the five health domains: clinical-epidemiological, normative, administrative, knowledge management and shared knowledge. The system proposes to contribute to the reduction of child mortality in Brazil. Thus, this paper presents studies over an intelligent module that uses Machine Learning to generate child death risk alerts on GISSA. These studies focus on trying different classification Algorithms, with a methodology based on Data Mining to reach a learning model capable of calculating the probability of a newborn dying. The work brings together public databases SIM and SINASC for the training of classification algorithms, identifying relationships between birth and death data of children under one year. During the methodological process, it was made a subsampling to balance the number of inputs and be fair in the training model results, executed with Matlab scripts.

Introduction

The problem of infant mortality mainly affects the so-called underdeveloped countries [21]. According to the United Nations, the overall rate of child mortality has dropped by 53% in 25 years [22], while in Brazil this reduction has been 77% in the last 22 years [23], probably due to improvements in maternal and child care through programs to support pregnant women, such as the Stork Network, whose goal is to preserve maternal and child health, especially in the first years of life [6]. The State of Ceará had a reduction of 11.5% between 2014 and 2015 [14]. However, these rates are still high compared with developed countries. Norway, for example, presented an infant mortality rate of 2.4% in 2014 [8]. Therefore, more effective strategies are needed to alleviate this problem.

GISSA (Intelligent Governance in Health Systems) is a framework to support decision making in health settings. The system is able to generate alerts and administrative reports for managers and health professionals.

This work presents the LAIS, a mechanism based on machine learning capable of predicting cases of infant mortality to assist managers in decision making. Integration and analysis of the public databases of the SIM (Mortality Information System) and SINASC (Information System for Live Births), made available by DATASUS (Department of Information Technology of the SUS) are made. Thus, the model generated by the LAIS is able, from the attributes of the newborn and those of its mother, to classify and calculate the risk of infant death.

This paper is organized as follows. Section 1 presents LARIISA and GISSA; Section 2 discusses related work; section 3 describes the intelligent module based on machine learning using the pattern recognition methodology and section 4 is presented to conclusions and future work.

1 Theoretical Foundation

1.1 LARIISA project

LARIISA is a platform developed in 2009 [20] with the aim of providing governance intelligence in the five areas of health (clinical and epidemiological, legal, administrative, knowledge management and shared knowledge) to help many users (patients, health workers, nurses, doctors, administrators, health secretaries, etc.) in decision making. To do so, it is necessary to manage health-related databases, dispersed on government bases or not, by crossing them with information captured in real time [15].

1.2 GISSA

The GISSA project is an instance of the LARIISA platform, with a focus on the Stork Network project of the Brazilian Ministry of Health, supported by FINEP (Funding of Studies and Projects), being implemented by Instituto Atlântico. Its purpose is to help decision makers, at all levels of the health cycle (patient, health agent, physician, hospital manager, secretary, etc.) by generating alerts and dashboards, available health bases, related to maternal and child health issues. A GISSA prototype is operational in Tauá, Ceará, and is being implemented in other municipalities of the State.

GISSA is therefore composed of a set of components that allows the collection, integration, and visualization of information relevant to the decision-making process [3]. Currently, it has the following alerts: live birth with low birth weight; delayed vaccination; prenatal; vaccine campaign; among others. In this context, [11] proposes a mechanism based on heuristics capable of calculating the probability of death of newborns using information from different databases for GISSA. However, although it is based on medical knowledge, the work does not perform tests of efficiency or precision, which prevents the evaluation of this mechanism.

2 Related Work

Malnutrition is considered to be a major cause of child mortality in underdeveloped countries. In [17], classification algorithms were used to find patterns related to the nutritional status of children under five years of age. The study aims to identify which factors affect the nutritional status of children. A total of 11,654 cases were treated with 16 health and socioeconomic attributes, collected from an Ethiopian Health Demographic Survey conducted in 2011. The machine learning algorithms used were J48 [25] of decision trees, Naive Bayes [16] and the rules inducer class PART [10]. After several experiments, were selected the PART algorithm that presented the best performance, with a precision of 92.6% and a Receiver Operating Characteristic (ROC) curve area of 97.8%.

In [28] a study on infant death in children under one year of age was performed using Data Mining techniques. Then, the SIM and SINASC databases were used for the municipality of Rio de Janeiro between 2008 and 2012. The integration was done through the field DN (Birth Certificate Number), present in the SINASC and SIM. A total of 3,336 individuals were born and died. In the research, the following 13 attributes were used: Sex of the Newborn, Apgar1 (5 parameters that are assessed during the first minute of the child's life - heart rate, respiration, muscle tone, irritability and skin color), Apgar5 parameters that are evaluated during the fifth minute of the child's life - heart rate, breathing, muscle tone, irritability and skin color), Newborn weight, Newborn color, Newborn age, basic cause of death , age of mother, number of dead children, number of live children, number of weeks of gestation, type of pregnancy and type of delivery. Was used the unsupervised algorithm Apriori [1] for the investigation of birth characteristics that are associated with death in children under one year of age. At the end of the work, some rules provided may assist health professionals.

A study of births at Bega Obstetrics and Gynecology Clinique, Timişoara, Romania, was presented in [27]. A data set was analyzed with 2,325 births and 15 attributes: mother's age, number of pregnancies, number of pregnancies weeks of gestation, child sex, child's weight, and type of delivery. The goal of the paper is to predict the child's Apgar score at birth, using the tool WEKA [9] and 10 classification algorithms: Naive Bayes, J48, IBK [2], Random Forest [5], SMO [24], AdaBoost [12], LogitBoost

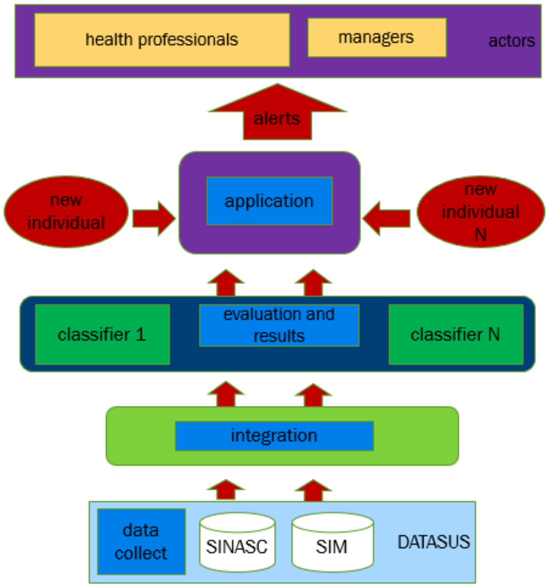

Figure 1: Smart Alert Generation Methodology

[13], JRipp [7], REPTree and SimpleCart [4]. The LogitBoost algorithm presented better results in the experiments. The generated model was used in a Java application to predict the Apgar score of a new patient. [18] uses Bayesian Networks to support decision making in uncertain environments. A network was developed to classify hypertensive disorders focused on the care of pre-eclampsia. Using the Bayesian Nisy-OR model in a database, the system analyzes the data layout and classifies them in the network. From the symptoms presented by the pregnant woman, the system, through statistical data, infers the severity of the case, helping the doctor who specializes in the diagnosis of pre-eclampsia. This approach proved accurate even with a small number of data.

[19] makes a detailed analysis between the Naive Bayes and the decision tree J48 classifiers. The paper analyzes a set of data related to hypertensive disorders to evaluate pregnancy complications. A study of the performance of the classifiers and the Confusion Matrix is done using predictive parameters. The two classifiers present close values. However, the J48 decision tree had a more accurate result.

3 Intelligent Module Based on Machine Learning

To achieve better results in the Data Mining process, the methodology of pattern recognition developed at the Federal University of Ceará (UFC) in the Centauro Laboratory was used. This methodology consists of a set of steps performed in the Data Mining process with the objective of selecting the best algorithms and attributes according to the context studied [26].

Figure 1 shows the steps developed in this work: Data collection; Integration; Evaluation and Results; and Application. First, data is collected from databases. Then, this data is integrated through the junction between the bases. Subsequently, the algorithms are trained, tested and evaluated according to the appropriate metric, generating a prediction model. Finally, the generated model is tested on a prototype capable of predicting the risk of a newborn coming to death.

3.1 Collection and Preparation of Data

Data were collected from two different public databases: SINASC, which contain information on live births; and SIM, which contain information on mortality, including cases of infant mortality. Both databases are available on the DATASUS portal in DBC (DataBase Container) format. The data refer to the State of Ceará in the years 2013 (SINASC and SIM) and 2014 (SIM). These data were converted to SQL (Structured Query Language) using TABWIN, a software provided by DATASUS for viewing and manipulating public data.

3.1.1 Integration and Selection of Attributes

With the relationship of SIM bases and SINASC it is possible to retrieve information about the birth of children victims of infant mortality. Thus it is possible to distinguish children who have survived or not up to one year of age.

Each live birth has a unique attribute called the Born Birth Declaration Number (numerodn), always filled in at the base of SINASC. The SIM base also has the field (numerodn), which is filled only in cases of infant death. This field was essential for the integration of the bases, since from it it is possible to relate the infant mortality data to the birth data. The integration was divided into 4 stages:

Step 01: Taking into account that children born in 2013 can be victims of infant mortality in 2014, the bases of SIM2013 and SIM2014 were united for children who died less than 1 year old. The following is a simplified expression (in relational algebra) of the integration process (Equation 1):

$$SIM' \leftarrow \sigma idade((SIM2013) \cup (SIM2014)) \tag{1}$$

Step 02: Then, the SINASC2013 and SIM are joined together using the numerodn field. The result returns all child mortality cases occurring in 2013 or 2014. The following is a simplified expression in relational algebra (Equation 2):

$$M \leftarrow (\rho SN(SINASC) \bowtie SN.numerodn = S.numerodn \rho S(SIM')) \tag{2}$$

Step 03: We also looked for cases of newborns who did not suffer death. Thus, a query was made at SINASC2013, except for cases that suffered infant death M. The following is a simplified expression in relational algebra (Equation 3):

$$V \leftarrow (SINASC2013 - M)) \tag{3}$$

Step 04: Finally, a union of the death cases M and non-infantile V deaths occurred in 2013 and 2014. The following is a simplified expression (in relational algebra) (Equation 4):

$$ALL \leftarrow (M \cup V) \tag{4}$$

In this stage, case-signaling was also performed, death YES and non-death NO, needed in supervised classification problems.

The result of the integration is a dataset with 50 attributes, containing information on the birth and death (if it occurred) of children born in 2013. The dataset obtained resulted in 1,182 cases of death and 124,876 cases of children who survived up to one year. 16 of these attributes were selected for the analysis step.

The values of these attributes are originally inserted as strings. The Matlab scripts used work with numbers, so it was necessary to convert it into to numeric values to make calculations. The weight for each category value was determined by the sequence it was observed in the data analysis.

The values were defined from -1 to 9, where:

- `-1` – used to ”Campo-em-Branco” (free translation: Blank field)
- `-0.5` – to ”Ignorado” (free translation: ignored)
- `0` – to ”Errado (free translation: wrong) and `from 1 to 9` – defined to the other characteristics considering from 1 to less death risk to 9 to bigger risk of death

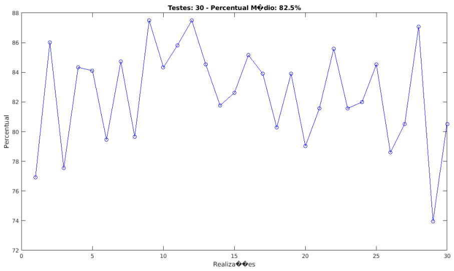

Figure 2: Classification with CrossValidation - MLP and KFold = 5

3.2 Analysis and Tests

In order to find a more adequate model for the prediction problem of infants, were performed experiments with the algorithms, K Nearest Neighbor (KNN), Naive Bayes (NB), Vector Support (SVM), Artificial Neural Networks (ANN), using Matlab scripts. Some of these algorithms are described as follows:

(i) K Nearest Neighbor (KNN): It works in order to calculate the similarity between the record to be analyzed and the records of the data set in order to estimate the class of the record that was presented as input. When a new record should be classified, it is compared to all training data records to identify k-neighbors closest according to some selected metric, and one of the most used is the Euclidean distance (Equation 5). The Euclidean distance refers to the distance between two points measured by the straight line that interconnects these two points (Equation 5).

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \tag{5}$$

(ii) Artificial Neural Networks: The letter in equation 6 represents the input signals (data on the problem), while the synaptic weights are represented by w_i (responsible for weighting the input signals according to the level of importance), and Σ represents the aggregating function. It also has + Θ, threshold of activation, a constant responsible for allowing or not the passage of signal; u represents the activation potential [29].

$$u = \sum_{i=1}^{n} w_i * x_i + \Theta \tag{6}$$

A cross-validation K-Fold was adopted because the stratification allows tests using different parameters, and determine the best learning rate and neurons number.

(iii) Naive Bayes: The Bayesian Naive Bayes algorithm is based on probability theory and assumes that attributes will influence the class independently. During model creation, the classifier will construct a table showing how much each category of each attribute contributes to each class. In equation 7, C represents the class and e={A1 = a1,...{An = an} are the attributes of the classes. The tests show that the Naive Bayes classifier is the most suitable for this purpose, presenting good results with area of the ROC curve of 92.1%.

$$P(A_1, ...A_n, C = P(C) \prod_{k=1}^{n} P(A_i|C) \tag{7}$$

3.3 Evaluation and Results

In the first experiment, the Naive Bayes algorithm obtained the best results. It presented the highest area value of the ROC curve compared to the other.

Table 1 refers to the experiment with balanced data, in which the Spread Subsample algorithm was used to balance the classes of the data sample by means of a random sub-sampling. This equals the number of individuals in the living and dead classes with, respectively, with 1,182 instances for each class. It is noticed that even with the change in the number of classes, the Naive Bayes algorithm continued to have the best result because it had the largest area value of the ROC curve.

Algorithms	Precision	Recall	F-Measure	ROC
KNN	0,895	0,830	0.861	0,888
Naive Bayes	0,921	0,809	0,861	0,924
V. PERCEPTRON	0,900	0,838	0,868	0,875
MLPerpectron	0,837	0,821	0,829	0,898

Table 1: Experiment

Because the Naive Bayes algorithm obtained the largest area value of the ROC curve in the experiment, a more detailed study was carried out in this experiment: Table 2 shows the Confusion Matrix of the Naive Bayes algorithm for an analysis of the results. It is verified that Naive Bayes correctly classified 2,056 children (86,912%) that correspond to the correct diagonal in the table below (956 + 1,100). Therefore, 308 children (13,028%) were classified incorrectly (another diagonal: 82 + 226).

		Predicted Class	
		Morto	Vivo
Real Class	Morto	956	226
	Vivo	82	1100

Table 2: Confusion Matrix - Naive Bayes

Among the 308 children who were misclassified, 82 (3.46%) were false positives and 226 (9.56%) were false negatives. Of the 2,056 children who were correctly classified, 956 (40.44%) are true positives and 1,100 (46.53%) are true negatives. As 956 are true positives, this indicates those who suffered childhood death and that 82 false positives did not suffer infant death, but were classified as having died.

3.4 Application

After a process of analysis and comparison between the algorithms, we made the choice of the most efficient classification algorithm according to the domain studied. Thus, the Naive Bayes classifier is the one that best adapts to the data set analyzed. Then, the model generated by the algorithm was used to classify the risk of new patients suffering death.

4 Conclusions and Future Work

The proposal presented in this paper adds value to the GISSA alerts, providing them with an intelligent mechanism based on classifiers. Thus, it is able to provide the health manager, in addition to the important warnings that already produced the probability of death of a newborn from the information of the pregnant and the newborn itself. Therefore, the decision maker may prioritize more urgent cases and, consequently, mitigate the serious problem of infant mortality.

As a future work, it is intended to apply the methodology used in the present work to the integration of SINASC and e-SUS databases, as also run tests with other tools like the language R and Scikit-python to test the perfomance of the tools itself. It is also expected to use together classification and heuristics with ontology (that is under work by other team) to fit specific classes of problems. This will allow the possibility of developing a hybrid mechanism to be added to the GISSA, from these experiments.

Acknowledgment

The authors would like to thank PRPI / IFCE and FUNCAP for the support they received via the Program for Productivity in Research, Incentives for Interiorization and Technological Innovation.

References

[1] Rakesh Agrawal, Ramakrishnan Srikant, et al. Fast algorithms for mining association rules. In *Proc. 20th int. conf. very large data bases (VLDB), September 12-15, Santiago, Chile*, volume 1215, pages 487–499, 1994.

[2] David W. Aha, Dennis Kibler, and Marc K. Albert. Instance-based learning algorithms. *Machine Learning*, 6(1):37–66, 1991.

[3] L. O. M. Andrade, M. Oliveira, and Ronaldo Ramos. Projeto GISSA: META FÍSICA 3 – atividade 3.1 Definir modelo de inteligência de gestão na saúde. https://amauroboliveira.files.wordpress.com/2015/11/2015-nov30-meta-3-ativ-1-moldelointeligc3aanciagestc3a3o-draf-1-0.pdf, 2015. [Online; accessed 30-September-2016].

[4] L Breiman, JH Friedman, RA Olshen, and CJ Stone. Classification and regression trees, wadsworth international group, belmont, CA, 1984. *Case Description Feature Subset Correct Missed FA Misclass*, 1:1–3, 1993.

[5] Leo Breiman. Random forests. *Mach. Learn.*, 45(1):5–32, October 2001.

[6] Pauline Cristine da Silva Cavalcanti, Garibaldi Dantas Gurgel Junior, Ana Ribeiro de Vasconcelos, and Andre copyright Vinicius Pires Guerrero. Um modelo da Rede Cegonha, 12 2013.

[7] William W Cohen. Fast effective rule induction. In *Proceedings of the twelfth international conference on machine learning, July 9—12, Tahoe City, CA, USA*, pages 115–123, 1995.

[8] CIA World Factbook. Noruega taxa de mortalidade infantil. http://www.indexmundi.com/pt/noruega/taxa_de_mortalidade_infantil.html/, last viewed: July 17 2017, 2015.

[9] Eibe Frank, Mark. Hall, and Ian Witten. Online appendix for ”data mining: Practical machine learning tools and techniques. In *Morgan Kaufmann*. 5 edition, 2016.

[10] Eibe Frank and Ian H. Witten. Generating accurate rule sets without global optimization. In *Proceedings of the Fifteenth International Conference on Machine Learning (ICML '98)*, pages 144–151, San Francisco, CA, USA, 1998. Morgan Kaufmann Publishers Inc.

[11] Renato Freitas, Cleilton Lima, Oton Braga, Gabriel Lopes, Odorico Monteiro, and Mauro Oliveira. Using linked data in the integration of data for maternal and infant death risk of the sus in the gissa project. In *Proceedings of the 23nd Brazilian Symposium on Multimedia and the Web (WebMedia '17), October 17–20, Gramado, RS, Brazil.* ACM, 2017.

[12] Yoav Freund, Robert E Schapire, et al. Experiments with a new boosting algorithm. In *icml*, volume 96, pages 148–156, 1996.

[13] Jerome Friedman, Trevor Hastie, Robert Tibshirani, et al. Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). *The annals of statistics*, 28(2):337–407, 2000.

[14] G1. Taxa de mortalidade infantil no ceará. http://g1.globo.com/ceara/noticia/2016/08/ceara-reduz-mortalidade-infantil-materna-e-fetal, last viewed: PREENCHER, 2016.

[15] Leonardo M ”Gardini, Reinaldo Braga, Jose Bringel, Carina Oliveira, Rossana Andrade, Hervé Martin, Luiz OM Andrade, and Mauro” Oliveira. Clariisa, a context-aware framework based on geolocation for a health care governance system. In *IEEE 15th International Conference on e-Health Networking, Applications & Services (Healthcom), October 9-12, Lisbon, Portugal*, pages 334–339. IEEE, 2013.

[16] George H John and Pat Langley. Estimating continuous distributions in bayesian classifiers. In *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence, August 18-20, Montreal, QU, Canada*, pages 338–345. Morgan Kaufmann Publishers Inc., 1995.

[17] Z Markos, F Doyore, M Yifiru, and J Haidar. Predicting under nutrition status of under-five children using data mining techniques: The case of 2011 ethiopian demographic and health survey. *J Health Med Inform*, 5:152, 2014.

[18] M. W. L. Moreira, J. J. P. C. Rodrigues, A. M. B. Oliveira, R. F. Ramos, and K. Saleem. A preeclampsia diagnosis approach using bayesian networks. In *2016 IEEE International Conference on Communications (ICC)*, pages 1–5, May 2016.

[19] M. W. L. Moreira, J. J. P. C. Rodrigues, A. M. B. Oliveira, K. Saleem, and A. Neto. Performance evaluation of predictive classifiers for pregnancy care. In *2016 IEEE Global Communications Conference (GLOBECOM)*, pages 1–6, Dec 2016.

[20] Mauro Oliveira, Carlos Hairon, Odorico Andrade, Regis Moura, Claude Sicotte, J-L Denis, Stenio Fernandes, Jerome Gensel, Jose Bringel, and Hervé Martin. A context-aware framework for health care governance decision-making systems: A model based on the brazilian digital tv, 2010.

[21] ONU. Onu: Meta global de mortalidade infantil será atingida com atraso de 11 anos. http://www.bbc.com/portuguese/noticias/2014/09/140916_unicef_meta_mortalidade_infantil_rm, last viewed: July 22 2017, 2014.

[22] ONU. Onu afirma que taxa de mortalidade infantil no mundo caiu pela metade em 25 anos. http://www.uai.com.br/app/noticia/saude/2015/09/09/noticias-saude,187094/onu-a\begingroup\let\relax\relax\endgroup[Pleaseinsert\PrerenderUnicode{}intopreamble]rma-que-taxa-de-mortalidade-infantil-no-mundo-caiu-pela-metade, last viewed: July 17 2017, 2015.

[23] ONU. Taxa de mortalidade infantil no brasil cai 77 https://istoe.com.br/324257_TAXA+DE+MORTALIDADE+INFANTIL+NO+BRASIL+CAI+77+EM+22+ANOS+DIZ+ONU/, last viewed: June 28 2017, 2015.

[24] John C. Platt. Advances in kernel methods. chapter Fast Training of Support Vector Machines Using Sequential Minimal Optimization, pages 185–208. MIT Press, Cambridge, MA, USA, 1999.

[25] J Ross Quinlan. *C4.5: programs for machine learning*. Elsevier, 2014.

[26] Ronaldo F. Ramos, César L. C. Mattos, Amauri H. Souza Júnior, Ajalmar R. Rocha Neto, Guilherme A. Barreto, Hélio A. Mazzal, and Márcio O. Mota. Heart diseases prediction using data from health assurance systems in models and methods for supporting decision-making in human health and environment protection. In *Nova Publishers, Nova York, NY, USA*. 2016.

[27] Raul Robu and Ştefan Holban. The analysis and classification of birth data. *Acta Polytechnica Hungarica*, 12(4), 2015.

[28] Cláudio Jesus Rosa. Aplicação de KDD nos dados dos sistemas SIM e SINASC em busca de padrões descritivos de óbito infantil no município do rio de janeiro, 2015.

[29] IN da Silva, Danilo Hernane Spatti, and Rogério Andrade Flauzino. Redes neurais artificiais para engenharia e ciências aplicadas. *São Paulo: Artliber*, pages 33–111, 2010.