Using Predictive Classifiers to Prevent Infant Mortality in the Brazilian Northeast

Ronaldo Ramos[1], Cristiano Sousa[1], Mário W. L. Moreira[1,2], Joel J. P. C. Rodrigues[2,3,4]
Mauro Oliveira[1], and Odorico Monteiro[5]

[1] Federal Institute of Science and Technology of Ceará (IFCE), Fortaleza, CE, Brazil
[2] Instituto de Telecomunicações, Universidade da Beira Interior, Covilhã, Portugal
[3] National Institute of Telecommunications (Inatel), Santa Rita do Sapucaí, MG, Brazil
[4] University of Fortaleza (UNIFOR), Fortaleza, CE, Brazil
[5]National Congress, Brasília, DF, Brazil

ronaldo@ifce.edu.br; {cristianocagece, amauroboliveira, odorico0811}@gmail.com; {mario.moreira, joeljr}@ieee.org

***Abstract*—Despite the fact that infant mortality rates have been decreased in recent years, this issue stills being considered alarming to Brazilian health system indicators. In this context, the GISSA framework, an intelligent governance framework for Brazilian health system, emerges as a smart system for the Federal Government program, called *Stork Network*. Its main objective is to improve the healthcare for pregnant women as well as their newborns. This application aims to generate alerts focusing on the health status verification of newborns and pregnant woman to support decision-makers in preventive actions that may mitigate severe problems. Therefore, this paper presents the LAIS, an Intelligent health analysis system that uses data mining (DM) to generate newborns death risk alerts through probability-based methods. Results show that the Naïve Bayes classifier presents better performance than the other DM approaches to the used pregnancy data set analysis of this work. This approach performed an accuracy of 0.982 and a Receiver Operating Characteristic (ROC) Area of 0.921. Both indicators suggest the proposed model may contribute to the reduction of maternal and fetal deaths.**

***Index Terms*—Decision support systems; Data mining; Bayes methods; Infant mortality; Medical conditions; Pregnancy**

I. Introduction

Infant mortality is a problem that affects all countries, with a higher incidence in those socially underdeveloped. According to the United Nations (UN), the death rate in Brazil fell 77% in 22 years [1]. Despite of this reduction, that rate is still considered very high. With the recent advances in information technology, much has been done to assist health systems managers in decision-making processes [2], [3], which are the use of intelligent solutions. For example, the use of DM techniques can make the system able of issuing warnings about a newborn risk of death. This is what was made in this work.

GISSA represents a framework developed from LARIISA [4], which is a governance decision-making support system for heathcare environments. In fact, GISSA is an instance of LARIISA, made for a Health Ministry program, named Stork Network, whose objective is to preserve the pregnant women's and fetus' health, especially in the first year of the newborns live. A GISSA prototype was implemented in the city of Tauá, in the state of Ceará (Brazilian Northeast). Currently, it has the following functionalities: generation of risk alerts for live births with low weight alerts about delayed vaccination, prenatal cares, vaccination campaigns, among others. This paper presents LAIS, an analyzer that uses DM techniques to issue alerts to the government health systems. This DSS uses the SIM (data about infant mortality) and SINASC (live birth data) database systems, which belongs to DATASUS company (public data processing company), and provides a predictive model capable of detecting future cases of infant mortality, enabling decision makers to do something to mitigate the problem.

This paper is organized as follows. Section II presents the LARIISA platform, describing the importance of GISSA and the process of knowledge discovery in databases (KDD). Section III shows some related work in this regards. Section IV describes the performed studies, the used DM and machine learning (ML) algorithms. Section V shows the developed analyzer for health alerts. Finally, in Section VI, it is discussed the importance of this work that brings real intelligence to the GISSA project and suggestion for further works.

II. Related Work

A. LARIISA

LARIISA is a platform that aims to provide intelligence to DSSs in health governance. Its data sources are the health-related and geographically dispersed databases [5]. An application scenario of LARIISA performs the following steps. The health data is captured by sensors, and actions are taken from the inference about these data, which can result in the sending of an ambulance or a health agent, purchase of medication, relocation of health agents, among others.

B. GISSA

The GISSA framework, an intelligent health systems governance, is a solution created from LARIISA to build a DSS for the Brazilian Health Ministry in the context of the "Stork Network" project. Figure 1 shows the architecture

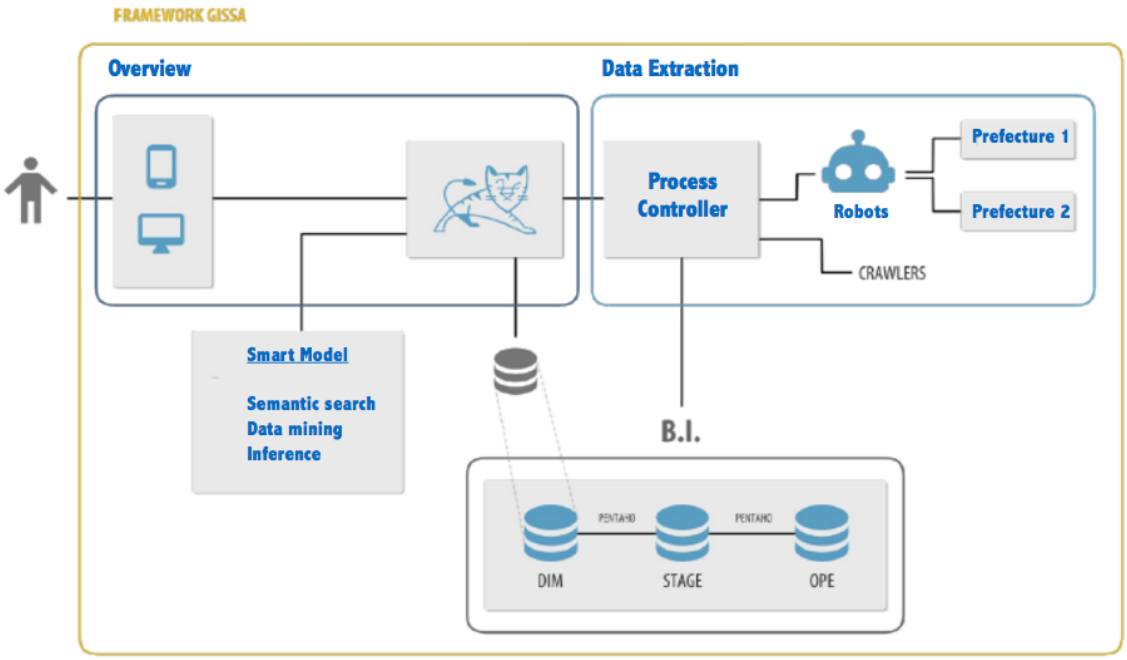

Fig. 1. GiSSA framework architecture [5].

of this framework. GISSA was used as a proof of concept (PoC) in the municipality of Tauá, CE, Brazil. The framework GISSA consists of a set of components that allow to collect, integrate and visualize relevant information to the decision-making process [6]. Currently, GISSA has alert systems to low weight newborns, delayed vaccination, prenatal care, and vaccine campaign, among others. Figure 2 shows the GISSA user interface.

Fig. 2. Tab of GISSA Alerts.

C. Other related works

In Markos *et al.*, classification algorithms were used to find patterns in the nutritional status of children under five years of age, considering that malnutrition is one of the main causes of infant mortality in underdeveloped countries [7]. The data used in this study were related to the 2011 Demographic Health Census of Ethiopia, which is conducted every five years. The aim of the study was to verify what attributes values affect the nutritional status of the children.

The ML algorithms used in that work were the ID3 [8] for decision tree construction, the Naïve Bayes [9] and the rule induction classifier PART [10]. The dataset used had 11 thousand instances and the features used were basically: weight, age, and height for children, and age, schooling, wealth index, address, the number of children, body mass index, occupation, for mothers. Besides, the size of the child at birth, vaccines were taken, child anemia level, gender, and nutritional status. After several experiments, the PART algorithm presented the best performance with a precision of 92.6% and area under the Receiver Operating Characteristic (ROC) curve 0.978.

A study about infant mortality with children under one-year old was performed in [11] using DM techniques over the integrated SIM and SINASC databases of the municipality of Rio de Janeiro, RJ, Brazil, between the years 2008 and 2012. A lot of work has been done to integrate these databases. It was possible to recover the data of 3336 individuals who were born and suffered infant death. This research used 13 features from the dataset instances such as: gender of the newborn, 1-minute Apgar score, this indicator refers to 5 parameters that are evaluated during the first minute of the child's life, to know, heart rate, breathing, muscle tone, irritability, and the color of the skin. Other instances are 5-minutes Apgar score, weight,

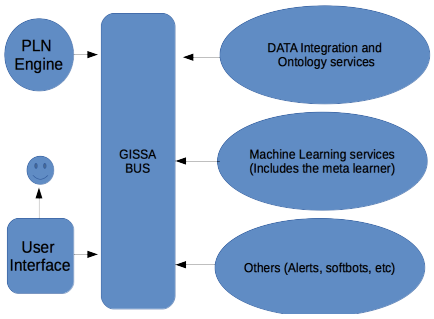

Fig. 3. Gissa Smart Model Components

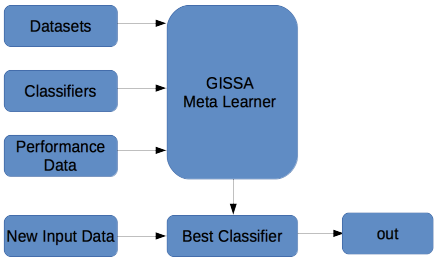

Fig. 4. Gissa Meta Learner

color, age of the child, the cause of the death, mother's age, number of mother's dead children, number of mother's living children, number of weeks of gestation, type of pregnancy, and type of birth. Good results were obtained in this research by the application of the non-supervised algorithm *a priori* whose objective was to investigate the birth characteristics that are associated with death in children under one year old in three different study scenarios [12]. As an important achievement, some rules were found that could help health professionals in their everyday activities.

A study on births in Bega Obstetrics and Gynecology Clinique, in Timisoara, Romania, was presented by Robu and Holban in 2010 [13]. It analyzed 2325 births based on 15 features such as mother's age, the number of gestations, number of weeks of pregnancy, child's gender, child's weight, and type of delivery. They looked for a new way to estimate the child's Apgar score at birth. To do so, they used 10 classification algorithms such as Naïve Bayes, ID3, KNN [14], Random Forest [15], SMO [16], AdaBoost [17], LogitBoost [18], JRipp [19], REPTree, and SimpleCart [20]. After several experiments, the LogitBoost algorithm was selected as the best algorithm among those mentioned above, and a Java application was created based on that.

III. Intelligence in GISSA framework

As we can see in figure 3 GISSA framework uses several smart computing tools is a service oriented approach. They vary from the mechanisms of ontologies, which are used primarily for the integration of databases, to the natural language processing engine, and the meta-learner. The meta-learner consists in the implementation of a strategy of selection of learning models that best fit the selected datasets. Next subsection discusses this element of the architecture.

A. GISSA Meta Learner

This system follows the work previously performed by [21] whose purpose is to select the best ML model to the context. In figure 4 we can see the Gissa meta learner architecture. The steps of the meta learner work are described below.

1) Performance evaluation of Selected models;
2) Data preparation;
3) Iterative construction of models;
4) Selection of The best approach;
5) Setup the correct model for production.

Next subsection discuss these steps.

1) Performance evaluation of Selected models: Selection of algorithms to evaluate and their hyperparameters. Sometimes, this work uses over a hundred of ML algorithms in search of the best. This research initially separated the models into groups, trying to use at least one algorithm of each except for support vector machines (SVM) that were not considered originally. The groups would therefore be:

- Decision Tree Algorithms: ID3, C4.5, Random Forest (RF), among others.
- Algorithms Based on Bayes' Theory: Bayes Net (BN) e Naïve Bayes (NB).
- Neural Networks: Voted Perceptron [22] and Multi Layer Perceptron (MLP).
- Kernell Methods: SVM.
- Elementary Classifiers: NN, KNN, CMD, etc.
- Rule Based: PART.

The use of homogeneous and heterogeneous ensembles is underway at the moment.

2) Data preparation: This work used data from the SIM and SINASC databases available on the DATASUS portal. For example, the practitioner can see in the Table I and II, respectively, the number of infant deaths in the state of Ceará for the years 2013 and 2014 and the number of live births in Ceará state into the year 2013. Table III shows the classes used in this study to predict the risk of fetal death.

It is important to mention that the system keeps track of whether or not it is a case of infant mortality. It is also part of this step to carry out an analysis of data completeness, an indication of data quality [23], as showed in Table IV.

As shown in Table IV, in the year of 2013 were registered 124, 876 births in the state of Ceará. Sixteen attribute features were selected in this study to construct the inference models.

TABLE I
SIM Database for 2013 and 2014.

SIM database	
Year	Number of Deaths
2013-2014	1.681

TABLE II
SINASC DATABASE FOR 2013.

SINASC database	
Year	Number of Births
2013	124.876

TABLE III
FEATURES EXTRACTED FROM SIM AND SINASC DATABASES TO THE SYSTEM MODELING.

#	**Feature**	**Description**
1	Age	Mother's age
2	Marital Status	Marital Status of the Mother
3	Schooling	Mother's Schooling
4	Localization	Birthplace
5	Number of live births	Number of live births in previous pregnancies
6	Number of children born dead	Number of children born dead in previous pregnancies
7	Gestation week	Number of weeks of gestation
8	Pregnancy	kind of pregnancy
9	Birth	Type of birth
10	Gender	Child's Gender
11	Weight	Infant weight at birth
12	Consultations	Prenatal consultations.
13	Apgar1	Apgar 1 minute
14	Apgar5	Apgar 5 minutes
15	Anomaly	Child born with anomaly
16	Color	Child skin color

Regarding the completeness of the data, the result is a median 99.58%, and eight attributes (50%) present excellent values. It can be seen that only "dead-born" (81.72%) and "live birth" (87.92%) attributes were below that (90%). Regarding the percentage of ignored data, we have "number of consultations" (1.77%) while all remaining attributes kept this percentage below 1%, so the data was considered as good quality. It is also important to mention that some downsampling had to be made to maintain the equilibrium among the database instances.

3) Iterative construction of the models: After selecting the models and preparing the data to be used, we carry out the training of the models. This process is done in sequence and individual results are recorded for comparison. Let us look at some of them.

TABLE IV
THE LEVEL OF DATA COMPLETENESS. THIS INDICATOR VERIFIES THAT THE ERRORS IN THE DATA ARE WITHIN ACCEPTABLE LIMITS.

Level of the Features Completeness		
Feature	Completeness (%)	Ignored (%)
Gender	100	0,0
Marital Status	98,7	0,32
Pregnancy	99,8	0,0
Age	100	0,0
Number of live births	87,923	0,0
Number of children born dead	81,72	0,0
Gestation week	93,18	0,0
Schooling	95,81	0,48
Consultations	100	1,77
Birth	99,75	0,0
Apgar1	99,41	0,02
Apgar5	99,41	0,02
Localization	100	0,0
Anomaly	94,69	0,25
Weight	100	0
Type of Pregnancy	99,8	0

TABLE V
PERFORMANCE EVALUATION OF THE PROPOSED METHODS.

Algorithm	PREC.	RECALL	F-MEAS.	AUROC
ID3	0,671	0,292	0,409	0,808
RF	0,64	0,289	0,399	0,883
BN	0,294	0,607	0,396	0,922
NB	0,294	0,607	0,396	0,921
KNN	0,479	0,273	0,348	0,785
Voted Perceptron	0,695	0,285	0,404	0,642
MLP	0,689	0,287	0,405	0,911
PART	0,567	0,306	0,398	0,857

TABLE VI
NB CONFUSION MATRIX.

		Predict	
		Dead	Alive
Actual	Dead	718	464
	Alive	1723	121971

As can be seen in the Table V and in figure 5, the NB and BN algorithms obtained better results during a specific experiment. Both presented the best recall value and area under ROC curve. A higher recall value will indicate a larger number of correctly classified samples as deaths over total death. With respect to the area under the ROC curve, when comparing classifiers using this metric, it is considered as the best the one that shows its closest to 1. This work performed the cross validation method to validate the proposed models.

4) Selection of the best approach: This architecture always points the best approach to a certain dataset. In a good number of cases, the MLP is usually the best, but the NB presented best results. One of the aspects that contributed to the selection of the NB algorithm was the ability of Bayesian algorithms to deal with incomplete and imprecise information [24]. Such performance may have been because it is probabilistic classifier based on the Bayes' theorem and assumes that the attributes will influence the class independently. The table VI shows the confusion matrix of the NB algorithm for a more detailed analysis of the results. It is noticeable that NB classified correctly $122,689$ cases (98.2487%) that correspond to the correct diagonal of Table VI. Therefore, $2,187$ (1.7513%) were incorrectly classified. Among the $2,187$ that were misclassified, 1723 (1.38%) are false positives and 464 (0.36%) are false negatives. From the $122,689$ that were correctly classified, 718 (0.57 %) are true positives and $121,971$ (97.67%) are true negatives. As 718 cases are true positives, this indicates those who suffered infant death and those $1,723$ false positives did not suffer infant death but were classified as patients who died.

5) Setup the correct model for production: The analysis of the results guides the selection of the most efficient classification algorithm for the dataset in question. After a thorough process of analysis and comparison of algorithms,

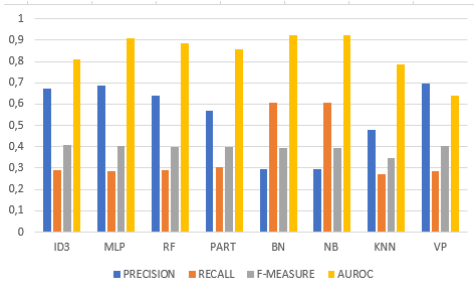

Fig. 5. Performance Evaluation of the Proposed Methods

using different approaches and strategies, it was concluded that the NB classifier is the one that best adapts to the data set analyzed. Therefore, it is time to code the solution for production use. More tests need to be done to deal with problems such as overfitting. After a final phase of tests and parameter adjustments, this research implements the application named LAIS. This system performs an inference mechanism with all the adjusted parameters. Uses some open source libraries like those made available by the software WEKA [25]. LAIS is a GISSA software service consisting of an interface, where information about mother and child is inserted; an intelligent model, which uses the NB classifier to calculate the probability of the occurring of an infant death. After entering all the information and clicking the Start button, the application captures the input data, applies a mathematical model, performs the classification, and shows the result in percentages on a screen.

IV. Conclusion and Future Work

This paper presents the GISSA framework that contains a series of software services with the purpose of assisting the process of decision making in health systems of government agencies. As a first case study, data science technologies were applied to build a predictive model for infant mortality for a northeastern region of Brazil that is exactly the region that is most sensitive to the problem. As a result of this study, an additional service to GISSA called LAIS was developed. LAIS is a software service capable of giving the probability of a child born in the NE region of Brazil to survive the first year of life. Together with other clinical data such as Apgar, the GISSA architecture, with its alerting system, can send these inferences to managers to take measures for each risky newborn in the region.

Further work suggests to apply the methodology used in the present work to an integrated view of SINASC and E-SUS data sources created by [26]. By this way, it will be possible to enrich the LAIS, identifying relationships among several factors of infant deaths and premature births with more information about mothers, such as alcohol, tobacco and/or drug use during pregnancy, among others.

Acknowledgments

This work was supported by the National Funding from the FCT - Fundação para a Ciência e a Tecnologia through the $UID/EEA/50008/2013$ Project, by Finep, with resources from Funttel, Grant No. 01.14.0231.00, under the Centro de Referência em Radiocomunicações - CRR project of the Instituto Nacional de Telecomunicações (Inatel), Brazil, by Ciência sem Fronteiras of CNPq, Brazil, through the process number $207706/2014-0$, and by Fundação Cearense de Apoio ao Desenvolvimento Científico e Tecnológico (FUNCAP) under the Program of Research Productivity Grants, Incentive for Interiorization and Technological Innovation (BPI), Edital No. 09/2015.

References

[1] A. Angulo-Tuesta, L. M. P. Santos, and D. A. Natalizi, "Impact of health research on advances in knowledge, research capacity-building and evidence-informed policies: a case study on maternal mortality and morbidity in brazil," *Sao Paulo Medical Journal*, vol. 134, no. 2, pp. 153–162, 2016.

[2] W. Moudani, M. Hussein, F. Mora-Camino *et al.*, "Heart disease diagnosis using fuzzy supervised learning based on dynamic reduced features," *International Journal of E-Health and Medical Communications (IJEHMC)*, vol. 5, no. 3, pp. 78–101, 2014.

[3] R. Veloso, F. Portela, M. F. Santos, J. Machado, A. da Silva Abelha, F. Rua, and Á. Silva, "Categorize readmitted patients in intensive medicine by means of clustering data mining," *International Journal of E-Health and Medical Communications (IJEHMC)*, vol. 8, no. 3, pp. 22–37, 2017.

[4] M. Oliveira, C. Hairon, O. Andrade, R. Moura, C. Sicotte, J. Denis, S. Fernandes, J. Gensel, J. Bringel, and H. Martin, "A context-aware framework for health care governance decision-making systems: A model based on the brazilian digital TV," in *IEEE International Symposium on a World of Wireless Mobile and Multimedia Networks (WoWMoM), Jun 4-17, Montreal, QC, Canada*. IEEE, 2010, pp. 1–6.

[5] L. M. "Gardini, R. Braga, J. Bringel, C. Oliveira, R. Andrade, H. Martin, L. O. Andrade, and M. 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*. IEEE, 2013, pp. 334–339.

[6] L. O. M. Andrade, M. Oliveira, and R. 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].

[7] 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*, vol. 5, p. 152, 2014.

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

[9] G. H. John and P. Langley, "Estimating continuous distributions in bayesian classifiers," in *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence, August 18-20, Montreal, QU, Canada*. Morgan Kaufmann Publishers Inc., 1995, pp. 338–345.

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

[11] C. J. Rosa, "Aplicao de KDD nos dados dos sistemas SIM e SINASC em busca de padres descritivos de bito infantil no municpio do rio de janeiro," 2015.

[12] R. Agrawal, R. Srikant *et al.*, "Fast algorithms for mining association rules," in *Proc. 20th int. conf. very large data bases (VLDB), September 12-15, Santiago, Chile*, vol. 1215, 1994, pp. 487–499.

[13] R. Robu and Ş. Holban, "The analysis and classification of birth data," *Acta Polytechnica Hungarica*, vol. 12, no. 4, 2015.

[14] D. W. Aha, D. Kibler, and M. K. Albert, "Instance-based learning algorithms," *Machine Learning*, vol. 6, no. 1, pp. 37–66, 1991. [Online]. Available: http://dx.doi.org/10.1007/BF00153759

[15] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, Oct. 2001. [Online]. Available: http://dx.doi.org/10.1023/A:1010933404324

[16] J. C. Platt, "Advances in kernel methods," B. Schölkopf, C. J. C. Burges, and A. J. Smola, Eds. Cambridge, MA, USA: MIT Press, 1999, ch. Fast Training of Support Vector Machines Using Sequential Minimal Optimization, pp. 185–208. [Online]. Available: http://dl.acm.org/citation.cfm?id=299094.299105

[17] Y. Freund, R. E. Schapire *et al.*, "Experiments with a new boosting algorithm," in *icml*, vol. 96, 1996, pp. 148–156.

[18] J. Friedman, T. Hastie, R. Tibshirani *et al.*, "Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors)," *The annals of statistics*, vol. 28, no. 2, pp. 337–407, 2000.

[19] W. W. Cohen, "Fast effective rule induction," in *Proceedings of the twelfth international conference on machine learning, July 9-12, Tahoe City, CA, USA*, 1995, pp. 115–123.

[20] L. Breiman, J. Friedman, R. Olshen, and C. Stone, "Classification and regression trees, wadsworth international group, belmont, CA, 1984," *Case Description Feature Subset Correct Missed FA Misclass*, vol. 1, pp. 1–3, 1993.

[21] R. F. Ramos, C. L. C. Mattos, A. H. S. Júnior, A. R. R. Neto, G. A. Barreto, H. A. Mazzal, and M. 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.

[22] Y. Freund and R. E. Schapire, "Large margin classification using the perceptron algorithm," *Machine Learning*, vol. 37, no. 3, pp. 277–296, 1999. [Online]. Available: http://dx.doi.org/10.1023/A:1007662407062

[23] R. R. German, L. Lee, J. Horan, R. Milstein, C. Pertowski, M. Waller *et al.*, "Updated guidelines for evaluating public health surveillance systems," *MMWR Recomm Rep*, vol. 50, no. 1-35, 2001.

[24] K. Faceli, A. C. Lorena, J. Gama, and A. Carvalho, "Inteligência artificial: Uma abordagem de aprendizado de máquina," *Rio de Janeiro: LTC*, vol. 2, p. 192, 2011.

[25] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2016.

[26] G. Lopes, V. Vidal, and M. Oliveira, "A framework for creation of linked data mashups: A case study on healthcare," in *Proceedings of the 22nd Brazilian Symposium on Multimedia and the Web (WebMedia '16), November 08–11, Teresina, PI, Brazil*. ACM, 2016, pp. 327–330.