LAÍS, um Analisador Baseado em Classificadores para a Geração de Alertas Inteligentes em Saúde

Cristiano Silva [1], Joyce Quintino[1], Oton C. Braga[1], Ronaldo Ramos[2], Odorico Monteiro[3], Mauro Oliveira[1]

[1]Instituto Federal de Educação, Ciência e Tecnologia do Ceará (IFCE)
Aracati, CE - Brasil

[2]Instituto Federal de Educação, Ciência e Tecnologia do Ceará (IFCE)
Fortaleza, CE - Brasil

[3]Universidade Federal do Ceará (UFC)
Fortaleza, CE - Brasil

{cristianocagece, joycequintino11, otoncbraga, ronaldo.ramos,
odorico0811, amauroboliveira}@gmail.com

Abstract. *Although the infant mortality index has been reduced in recent years, this issue is still considered a serious problem in Brazilian health system indicators. In this context, the GISSA Framework (Intelligent Governance Framework for Brazilian Health System) emerges as a framework for the Federal Government program, called "Rede Cegonha". The main objective is to improve the health care for pregnant woman as well as the newborn. This framework aims to generate alerts focusing on the health status verification of newborns and pregnant woman in order to help healthy decision-makers in preventive actions that may mitigate the problem. Therefore, this paper presents the LAIS, an Intelligent Health System Analyzer based on data mining classifiers, which the objective is to generate alerts. Finally, we present the proposal results of an application that provides the death probability of a newborn, based on the analysis of his attributes and his mother.*

Resumo. *Embora nos últimos anos o índice de mortalidade infantil tenha sido reduzido, este tema ainda é considerado um grave problema nos indicativos da saúde no Brasil. O GISSA (Governança Inteligente em Sistema de Saúde) é um framework destinado ao Programa Rede Cegonha do Governo Federal, cujo objetivo é preservar a saúde da gestante e do recém-nascido. Este framework tem a função de gerar alertas relativos ao estado de saúde do recém-nascido e da gestante, de modo a ajudar os tomadores de decisão em saúde nas ações preventivas que possam mitigar o problema. Este trabalho propõe o LAÍS, um analisador baseado em mineração de dados, com objetivo de tornar inteligentes os alertas em Sistemas de Saúde. São apresentados os resultados de uma aplicação que fornece a probabilidade de um recém-nascido vir à óbito, a partir da análise de seus atributos e de sua mãe.*

1. Introdução

A mortalidade infantil é um problema que atinge todos os páıses, com maior incidência naqueles socialmente subdesenvolvida. De acordo com a Organização das Nações Unidas (ONU), a taxa de mortalidade no Brasil caiu 77% em 22 anos [ONU 2016]. Embora em redução no Brasil, esta taxa ainda é considerada muito elevada.

Com o avanço na tecnologia da informação, muito tem sido feito para auxiliar os gestores de saúde nos processos de tomada de decisão. Ela oferece meios que podem melhorá-los a partir da utilização de soluções inteligentes. Por exemplo, o uso de técnicas de mineração de dados pode tornar o sistema capaz de emitir alertas sobre o risco que um recém-nascido possui de vir a óbito. É o que se propõe neste trabalho.

O GISSA é um *framework* desenvolvido a partir do LARIISA [Oliveira et al. 2010], um sistema inteligente de governança para o apoio à tomada de decisão em ambientes de saúde. Assim, o GISSA é uma instância do LARIISA destinado ao Programa Rede Cegonha do Ministério da Saúde, cujo o objetivo é preservar a saúde da mãe e da criança, em especial nos primeiros anos de vida. Um protótipo do GISSA está sendo implementado na cidade de Tauá, no Ceará. Atualmente, ele dispõe das seguintes funcionalidades: geração de alertas de um nascido vivo com baixo peso, vacinação atrasada, pré-natal, campanha de vacina, entre outras. Contudo, essas funcionalidades ainda não fazem uso de mecanismos inteligentes.

Este trabalho apresenta o LAÍS, um analisador que utiliza técnicas de Mineração de Dados, para emissão de alertas para sistemas de saúde. Na prototipação foram usadas como dados as bases dos sistemas SIM (dados sobre mortalidade) e do SINASC (dados sobre nascidos vivos), ambos pertencentes ao DATASUS (Departamento de Informática do SUS). O resultado foi um modelo de previsão capaz de classificar novos casos de mortalidade infantil, permitindo a tomadores de decisão de mitigar o problema da mortalidade infantil, a partir de dados do recém-nascido e de sua mãe.

Este trabalho está organizado da seguinte forma. Na seção 2 é apresentado o LARIISA, descrevendo a importância do GISSA e do processo de Descoberta de Conhecimento em Bases de Dados, além de se discutir o conceito de Mineração de Dados para o propósito buscado; na seção 3, são abordados trabalhos relacionados ao presente contexto; na seção 4, são descritos o estudos realizados, a tarefa de Mineração de Dados usada, os algoritmos de Aprendizagem de Máquina; na seção 5, é apresentado o LAÍS o analisador desenvolvido para alertas em saúde; na seção 6, é discutida a importância desse trabalho que agrega inteligência aos alertas do projeto GISSA.

2. Fundamentação Teórica

2.1. LARIISA

O LARIISA é uma plataforma que visa prover inteligência de governança na tomada de decisão em sistemas de saúde, a partir do manejo de bases de dados relacionados à saúde, dispersos em bases governamentais ou não, cruzando-as com informações capturadas em tempo real [Gardini et al. 2013].

A figura 1 mostra um cenário de aplicação do LARIISA: dados de saúde são capturados por sensores, e ações são tomadas a partir da inferência sobre esses dados, podendo

resultar no envio de uma ambulância ou de um agente de saúde, compra de medicamento, regulação em hospitais, remanejamento de agentes de saúde.

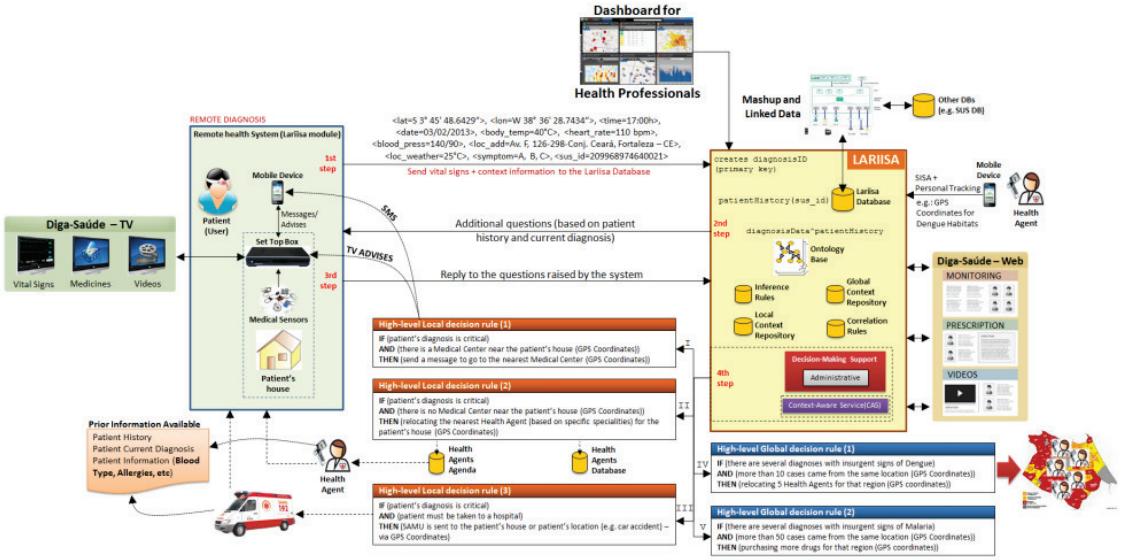

Figura 1. Cenário de internação domiciliar [Gardini et al. 2013].

2.2. GISSA

O *framework* GISSA (Governança Inteligente dos Sistemas de Saúde) é uma solução criada a partir do LARIISA para construção de sistemas de informação que apoiem o processo de tomada de decisão no contexto do projeto Rede Cegonha do Ministério da Saúde. O projeto GISSA implementou uma Prova de Conceito (PoC) no município de Tauá–CE. O *framework* GISSA é formado por um conjunto de componentes que permitem a coleta, integração e visualização de informações revelantes ao processo de tomada de decisão [Andrade et al. 2015].

Atualmente, o GISSA dispõe dos seguintes alertas: nascido vivo com baixo peso; vacinação atrasada; relacionadas ao pré-natal; campanha de vacina; entre outros (figura 2).

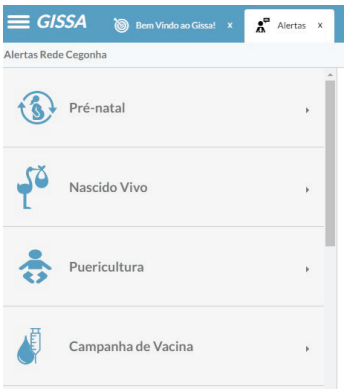

Figura 2. Alertas GISSA

2.3. Descoberta de Conhecimento em Bases de Dados

Em diversas aplicações, onde é necessário manipular uma grande quantidade de dados, o processo de descoberta de conhecimento em bases de dados (*Knowledge Discovery in Database-KDD*) tem o objetivo de extrair novas informações desses dados. FAYYAD define KDD como sendo "um processo interativo e iterativo, não trivial, constituído por diversas etapas, de extração de informações implícitas, anteriormente desconhecidas e potencialmente úteis, a partir dos dados armazenados" [FAYYAD et al. 1996]. O termo interativo refere-se a necessidade da atuação do homem como responsável pelo controle do processo, ou seja, analisar e interpretar os resultados obtidos ao longo do processo. Já o termo iterativo sugere a necessidade de repetições do processo de KDD, a fim de buscar os melhores resultados por meio de sucessivos refinamentos.

3. Trabalhos Relacionados

Em [Markos et al. 2014] foram utilizados algoritmos de classificação para encontrar padrões relativos ao estado nutricional de crianças menores de cinco anos, considerando-se que a desnutrição é um dos principais causadores de mortalidade infantil em países subdesenvolvidos. Os dados utilizados nesse estudo foram relativos à Pesquisa Demográfica de Saúde da Etiópia, em 2011, gerados em um censo realizado em intervalos de cinco anos. O estudo teve como objetivo verificar se os valores dos atributos afetam o estado nutricional das crianças. O software utilizado neste trabalho foi o WEKA [Frank et al. 2016] e algoritmos foram J48 [Quinlan 1993] de árvores de decisão, Naive Bayes [John and Langley 1995] e o classificador de indução de regras PART [Frank and Witten 1998]. Nesse trabalho foi criado um *data-set* com 11.654 instâncias e 16 atributos. Esses atributos são: peso da criança, idade da criança, altura da criança, idade da mãe, escolaridade da mãe, índice de riqueza da mãe, local da residência, número de crianças, índice de massa corpórea da mãe, ocupação da mãe, tamanho da criança ao nascer, vacinação, nível de anemia da criança, sexo da criança, idade da criança e estado nutricional. Após diversos experimentos foi selecionado o algoritmo PART, que apresentou o melhor desempenho tendo precisão de 92,6% e área da curva de ROC(*Receiver Operating Characteristic*) 97,8%.

Em [ROSA 2015] foi realizado um estudo sobre óbito infantil em crianças menores de um ano utilizando técnicas de Mineração de Dados, fazendo uso das bases de dados do SIM e SINASC integradas do Município do Rio de Janeiro entre os anos de 2008 a 2012. Para integrar essas duas bases de dados, usou-se o campo DN (Número de Nascimento), presente no SINASC e no SIM. Quando a criança sofre ó bito e tem idade menor do que um ano, esse campo é preenchido no SIM; quando ocorre o nascimento de uma criança esse campo é preenchido no SINASC. Assim, este campo permite relacionar os dados do SIM com os do SINASC. Depois da integração destes dados foi possível relacionar um total de 3336 indivíduos que nasceram e sofreram óbito infantil. Na pesquisa foram usados 13 atributos: sexo do RN (Recém-Nascido), Apgar1[1], Apgar5[2], peso, cor do RN, idade do RN, causa básica da morte, idade da mãe, quantidade de

[1]Refere-se a 5 parâmetros que são avaliados, durante o primeiro minuto de vida da criança, sendo esses frequência cardíaca, respiração, tônus muscular, irritabilidade e cor da pele

[2]Refere-se a 5 parâmetros que são avaliados, durante o quinto minuto de vida da criança, sendo esses frequência cardíaca, respiração, tônus muscular, irritabilidade e cor da pele

filhos mortos, quantidade de filhos vivos, número de semanas de gestação, tipo da gravidez e tipo do parto. Foi utilizado o algoritmo de aprendizado não supervisionado Apriori [Agrawal et al. 1994] a fim de investigar as caraterísticas de nascimento que estão associadas ao óbito em menores de um ano de idade e três cenários de estudo. Ao final do trabalho, foram encontradas algumas regras que podem auxiliar os profissionais de saúde.

Em [Robu and Holban 2015] foi apresentado um estudo sobre os nascimentos ocorridos no *Bega Obstetrics and Gynecology Clinique, Timişoara*, Romênia em 2010. Foi analisado um conjunto de dados com um total de 2.325 nascimentos, com base em 15 atributos tais como: idade da mãe, número de gestações, número de semanas de gestação, sexo da criança, peso da criança e tipo do parto. Buscou-se selecionar um algoritmo para prever a pontuação do Apgar da criança ao nascer. Para tanto, foram utilizados a ferramenta WEKA e 10 algoritmos de classificação sendo esses, Naive Bayes, J48, IBK [Aha et al. 1991], Random Forest [Breiman 2001], SMO [Platt 1999], AdaBoost [Freund et al. 1996], LogitBoost [Friedman et al. 2000], JRipp [Cohen 1995], REPTree e SimpleCart [Breiman et al. 1984]. Após vários experimentos selecionou-se o algoritmo LogitBoost como melhor algoritmos entre os citados anteriormente e criado uma aplicação em Java utilizando o modelo criado com algoritmo LogitBoost para prever a pontuação Apgar de um novo paciente.

4. Metodologia de Estudo

O estudo realizado neste trabalho seguiu a Metodologia de Reconhecimento de Padrões desenvolvida na UFC (Universidade Federal do Ceará) no Laboratório Centauro que consiste em um conjunto de passos (etapas) a serem desenvolvidos no processo de Mineração de Dados, cujo objetivo é fazer com que sejam selecionados os melhores algoritmos de acordo com o contexto estudado [Ramos et al. 2016].

4.1. Seleção inicial

Foi selecionado o WEKA, por se tratar de uma das ferramentas mais utilizadas no ambiente acadêmico:

- O WEKA ganhou o SIGKDD Data Mining que é o prêmio de descoberta de conhecimento [Piatetsky-Shapiro 2005].
- Licença do tipo GPL (General Public License).
- Multiplataformas Windows, Mac OS e Linux.
- Apresenta uma grande quantidade de algoritmos de classificação.
- Tem uma poderosa API que permite a integração em sistemas desenvolvidos em Java.
- Facilidade de uso por meio de sua interface gráfica.

4.2. Integração e Preparação dos Dados

As bases de dados utilizadas nesta pesquisa foram o SIM e o SINASC, disponíveis no portal do DATASUS[3]. Fez-se a integração dos dados por meio de consultas SQL*(Structured Query Language)*.

[3]http://www2.datasus.gov.br/DATASUS/index.php?area=0901item=1acao=28pad=31655

As tabela 1 e 2 mostram, respectivamente, a quantidade de óbitos infantis no estado do Ceará referentes aos anos de 2013 e 2014 e a quantidade de nascidos vivos no estado do Ceará referente ao ano de 2013.

Tabela 1. Dados do SIM referentes aos anos de 2013 a 2014

SIM	
ANO	NÚMERO DE ÓBITOS
2013-2014	1.681

Tabela 2. Dados do SINASC referente ao ano de 2013

SINASC	
ANO	NÚMERO DE NASCIMENTOS
2013	124.876

Os dados foram acessados através do TABWIN(TAB para WINdows), um *software* gratuito de tabulação disponível no site do DATASUS que também permite a conversão de arquivos dbc para dbf e deste último para o formato SQL.

Para a relação entre as bases, foi utilizado o atributo numerodn, um campo presente nas duas bases, desde que ocorra o óbito. Após a identificação do atributo capaz de relacionar as bases, foram feitas as consultas SQL. Em seguida, relacionou-se 1.182 indivíduos de um total de 1.681 que sofreram óbito.

Foi realizada uma análise com os dados do SIM e SINASC, observando os dados encontramos alguns campos não preenchidos limitando a quantidade de dados na pesquisa. De acordo com essa análise, foram selecionados 16 atributos: idade, estado civil, escolaridade, local de nascimento, quantidade de filhos vivos, quantidade de filhos mortos, gestação, gravidez, parto, sexo, peso, consultas, Apgar1, Apgar5, anomalia e cor (tabela 3).

Tabela 3. Atributos

Nº	Atributo	Resumo da descrição dos atributos
1	Idade	Idade da mãe
2	Estado civil	Estado civil da mãe
3	Escolaridade	Nível de escolaridade da mãe
4	Local	Local de nascimento da criança
5	Quantidade de filhos nascidos vivos	Número de filhos nascidos vivos nas gestações anteriores
6	Quantidade de filhos nascidos mortos	Número de filhos nascidos mortos nas gestações anteriores
7	Gestação	Número de semanas de gestação
8	Gravidez	Tipo da gravidez
9	Parto	Tipo do parto da mãe
10	Sexo	Sexo da criança
11	Peso	Peso da criança ao nascer
12	Consultas	Número de consultas pré-natal.
13	Apgar 1 minuto	Refere-se a 5 parâmetros que são avaliados, durante o primeiro minuto de vida da criança
14	Apgar 5 minutos	Refere-se a 5 parâmetros que são avaliados, durante o quinto minuto de vida da criança
15	Anomalia	Criança nascida com anomalia congênita
16	Cor	Cor da criança

A figura 3 mostra uma tabela de dados gerada a partir de buscas SQL nas bases do SIM e SINASC onde na última coluna é possível se identifica se o paciente foi a óbito por mortalidade infantil (YES) não (NO).

Por último, estes dados foram convertidos para o formato CSV (*Comma-Separated Values*) e, posteriormente, para o formato do padrão WEKA: ARFF (*Attribute-Relation File Format*).

parto character(1)	consultas character(1)	sexo character(1)	apgar1 character(2)	racacorn character(1)	apgar5 character(2)	locnasc character(1)	idanomal character(1)	peso character(4)	morto text
1	2	1	09	4	10	3	2	3500	NO
1	3	1	08	4	09	3	2	3000	NO
1	4	2	08	4	09	1	2	3010	NO
1	4	2	08	1	09	1	2	2940	NO
1	4	1	08	1	09	1	2	3180	NO
1	4	1	08	1	09	1	2	3620	NO
1	4	1	08	1	09	4	2	3250	NO
1	3	1	09	4	10	3	2	3650	NO
1	4	2	09	4	10	1	2	2920	NO
1	3	1	09	1	10	1	2	3100	NO
1	4	2	08	1	09	3	2	3800	NO

Figura 3. Tabela de dados

Após a geração da tabela (figura 3) foi realizada uma análise com o WEKA no âmbito de verificar o nível de completude dos dados, um indicativo da qualidade dos dados [German et al. 2001] (tabela 4).

Tabela 4. Nível de completude dos dados do SINASC 2013

Nível de completude dos atributos		
Atributo	Completude (%)	Ignorado (%)
Sexo	100	0,0
Estado cívil	98,7	0,32
Tipo da gravidez	99,8	0,0
Idade da Mãe	100	0,0
Filhos nascidos vivos	87,923	0,0
Filhos nascidos mortos	81,72	0,0
Quantidade de semanas de gestação	93,18	0,0
Escolaridade da mãe	95,81	0,48
Número de consultas da mãe	100	1,77
Tipo do parto	99,75	0,0
Apgar 1 minuto	99,41	0,02
Apgar 5 minutos	99,41	0,02
Local de nascimento	100	0,0
Anomalia	94,69	0,25
Peso	100	0
Tipo da gravidez	99,8	0

Como mostrado na tabela 4, no ano de 2013 foram registrados 124.876 nascimentos no estado do Ceará. Foram analisados os 16 atributos selecionados neste trabalho (tabela 3), presentes na Declaração de Nascidos Vivos (DNV). Quanto à completude em seu preenchimento, o resultado é uma mediana 99,58 % e 8 atributos (50 %) apresentam valor acima desse valor. Percebe-se que apenas os atributos "nascidos mortos" (81,72 %) e "nascidos vivos" (87,92 %) apresentaram valor abaixo de 90 %. Em relação ao percentual de dados ignorados tem-se "número de consultas da mãe" (1,77 %); todos os atributos restantes mantiveram um percentual ignorado abaixo de 1 %, logo os dados foram considerados de boa qualidade.

Atributos ignorados e em branco nos sistemas de monitoramento são causados por uma série de deficiências, falta de informação nos prontuários indo até o desconhecimento de certas informações pelos acompanhantes do paciente. Isso pode ser oriundo da falta de cuidado e da importância concedida ao preenchimento da DNV pelo profissional responsável [Costa and Frias 2009].

4.3. Análise e Testes

Nesta etapa, foram realizados vários experimentos com oito algoritmos de classificação do WEKA (tabela 5), adotando um *cross-validation* 10x, pois a estratificação reduz a variância estimada, além de evitar altos custos computacionais [Japkowicz and Shah 2011]. Depois disso, listou-se os resultados dos algoritmos obtidos durante o experimento, sendo esses:

- Algoritmos baseados em árvore de decisão : J48 e Random Forest.
- Algoritmos baseados na teoria Bayesiana: Bayes Net e Naive Bayes.
- Algoritmos baseados em redes neurais: Voted Perceptron [Freund and Schapire 1999] e MLP (Multi Layer Perceptron).

- Algoritmos baseados no vizinho mais próximo: IBK.
- Algoritmos baseados em regras: PART.

Foi observado, o problema do desequilíbrio de classes, pois o quantitativo de crianças que nasce é bastante superior ao quantitativo de crianças que morre antes de completar um ano de idade. Apesar das ocorrências infrequentes, classificação correta de uma classe rara(óbito) nesta situação possui uma importância maior do que a classificação correta da classe majoritária(vivo).

4.4. Avaliação e Resultados

Como pode ser observado na tabela 5, os algoritmos Naive Bayes e Bayes Net obtiveram melhores resultados durante essa fase. Ambos apresentaram um maior valor de *recall* e área da curva ROC do que os demais. Um maior valor de *recall* indicará uma quantidade maior de amostras classificadas corretamente como óbitos sobre o total de óbitos e em relação a área da curva ROC, ao se comparar classificadores utilizando essa métrica é selecionado como melhor aquele apresentar o valor da área da curva ROC o mais próximo de 1.

Tabela 5. Experimento

Algoritmos	PRECISION	RECALL	F-MEASURE	ÁREA ROC
J48	0,671	0,292	0,409	0,808
RANDOM FOREST	0,64	0,289	0,399	0,883
BAYES NET	0,294	0,607	0,396	0,922
NAIVE BAYES	0,294	0,607	0,396	0,921
IBK	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

Como pode ser observado na tabela 6, os algoritmos Naive Bayes e Net Bayes obtiveram resultados próximos durante a etapa de Mineração de Dados.

Tabela 6. Matriz de confusão algoritmo Naive Bayes

		Classe predita	
		Morto	Vivo
Classe Real	Morto	718	464
	Vivo	1723	121971

Um dos aspectos que contribuiu para a seleção do algoritmo Naive Bayes foi a caraterística dos algoritmos bayesianos de lidarem bem com informações incompletas e imprecisas [FACELI 2015]. Tal desempenho, pode sido ocasionado por ele ser um classificador probabilístico baseado no teorema de Bayes e supõe que os atributos vão influenciar a classe de modo independente. A tabela 6 mostra a matriz de confusão do algoritmo Naive Bayes, para uma análise mais detalhada dos resultados. Verifica-se que o

Naive Bayes classificou corretamente 122.689 (98,2487 %) que correspondem a diagonal de acerto da tabela 6 (718 + 121.971) e, portanto, 2.187 (1,7513 %) foram classificados incorretamente (outra diagonal: 1723 + 464).

Dentro dos 2.187 que foram classificados erroneamente, 1.723 (1,38 %) são falsos positivos e 464 (0,36 %) são falsos negativos. Dos 122.689 que foram classificados corretamente, 718 (0,57 %) são verdadeiros positivos e 121.971 (97,67 %) são verdadeiros negativos. Como 718 são verdadeiros positivos, isso indica os que sofreram óbito infantil e que 1.723 falsos positivos não sofreram óbito infantil, mas foram classificados como pacientes que sofreram óbito.

5. LAÍS, um Analisador para Casos de Risco de Óbito Infantil

A análise dos resultados guia a escolha do algoritmo de classificação mais eficiente para o caso em questão. Após um processo minucioso de análise e comparação de algoritmos, usando diversas abordagens e estratégias, pôde-se concluir que o classificador Naive Bayes é o que melhor se adapta ao conjunto de dados analisado.

Foi desenvolvida uma aplicação em Java utilizando uma API (*Application Programming Interface*) para Mineração de Dados disponibilizada pelo WEKA.

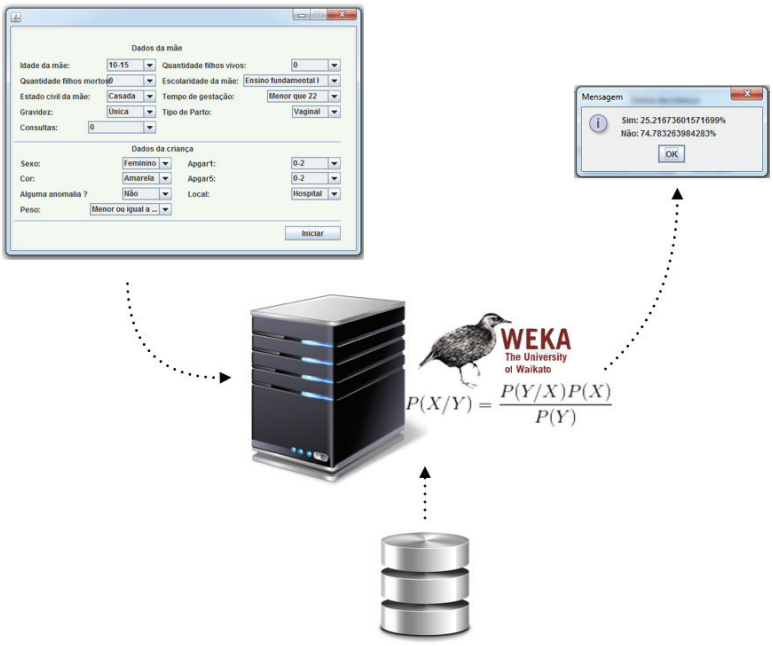

Figura 4. Arquitetura da aplicação

Trata-se de um protótipo (figura 4) inicial constituído de: (i) interface, onde são inseridas informações relativas a mãe e a criança; (ii) modelo inteligente, que usa o classificador Naive Bayes para calcular a probabilidade de ocorrer óbito infantil. Após inserir todas as informações e clicar no botão Iniciar, a aplicação captura os dados, gera um modelo matemático, realiza a classificação e mostra o resultado em percentuais numa tela.

6. Conclusões e Trabalhos Futuros

GISSA é um projeto FINEP (Financiadora de Estudos e Projetos), em execução pelo Instituto Atlântico, que ajuda tomadores de decisão, em todos os níveis do ciclo de saúde (paciente, agente de saúde, médico, gestor de hospital, secretário, etc.), mediante a geração de alertas, a partir da análise de dados nas diversas bases de saúde disponíveis. Este projeto tem apresentado excelentes resultados em sua prova de conceito no município de Tauá (Ce). A expectativa deste trabalho é agregar valor aos alertas GISSA. Por exemplo, o GISSA será capaz de fornecer ao gestor de saúde, além dos importantes alertas que já produzia, a probabilidade de óbito de um recém-nascido a partir das informações da gestante e, naturalmente, do próprio recém-nascido. A expectativa é de que o tomador de decisão possa, assim, priorizar casos mais urgentes e, consequentemente, mitigar o grave problema da mortalidade infantil.

Como trabalho futuro, pretende-se aplicar a metodologia utilizada no presente trabalho à uma visão integrada das fontes de dados SINASC e E-SUS criada por [Lopes et al. 2016]. Em assim procedendo, será possível enriquecer o LAÍS, identificando relações entre diversos fatores de óbitos infantis e partos prematuros com mais informações sobre as mães, tais como uso de álcool, tabaco e/ou drogas durante a gravidez, entre outras.

Referências

Agrawal, R., Srikant, R., et al. (1994). Fast algorithms for mining association rules. In *Proc. 20th int. conf. very large data bases, VLDB*, volume 1215, pages 487–499.

Aha, D. W., Kibler, D., and Albert, M. K. (1991). Instance-based learning algorithms. *Machine Learning*, 6(1):37–66.

Andrade, L. O. M., Oliveira, M., and Ramos, R. (2015). Projeto GISSA: Meta fÍsica 3 – atividade 3.1 definir modelo de inteligência de gestão na saúde.

Breiman, L. (2001). Random forests. *Mach. Learn.*, 45(1):5–32.

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984). Classification and regression trees belmont. *CA: Wadsworth International Group*.

Cohen, W. W. (1995). Fast effective rule induction. In *Proceedings of the twelfth international conference on machine learning*, pages 115–123.

Costa, J. M. B. d. S. and Frias, P. G. d. (2009). Avaliação da completitude das variáveis da declaração de nascido vivo de residentes em pernambuco, brasil, 1996 a 2005. *Cadernos de Saúde Pública*, 25(3):613–624.

FACELI, Katti; LORENA, A. C. G. J. C. D. C. A. (2015). *Inteligência Artificial: Uma Abordagem de Aprendizado de Máquina.* LTC, 1 edition.

FAYYAD, U., PIATETSKY-SHAPIRO, G., and SMYTH, P. (1996). Advances in knowledge discovery and data mining. In *American Association for Artificial Intelligence*.

Frank, E., Hall, M., and Witten, I. (2016). Online appendix for "data mining: Practical machine learning tools and techniques. In *Morgan Kaufmann*. 5 edition.

Frank, E. and Witten, I. H. (1998). Generating accurate rule sets without global optimization. *Machine learning*.

Freund, Y. and Schapire, R. E. (1999). Large margin classification using the perceptron algorithm. *Machine Learning*, 37(3):277–296.

Freund, Y., Schapire, R. E., et al. (1996). Experiments with a new boosting algorithm. In *icml*, volume 96, pages 148–156.

Friedman, J., Hastie, T., Tibshirani, R., et al. (2000). Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). *The annals of statistics*, 28(2):337–407.

Gardini, L. M., Braga, R., Bringel, J., Oliveira, C., Andrade, R., Martin, H., Andrade, L. O. M., and Oliveira, M. (2013). Clariisa , a context-aware framework based on geolocation for a health care governance system. *2013 IEEE 15th International Confeence on e-Health Networking, Applications and Services(Healthcom 2013)*, pages 334–339.

German, R. R., Lee, L., Horan, J., Milstein, R., Pertowski, C., Waller, M., et al. (2001). Updated guidelines for evaluating public health surveillance systems. *MMWR Recomm Rep*, 50(1-35).

Japkowicz, N. and Shah, M. (2011). *Evaluating learning algorithms: a classification perspective*. Cambridge University Press.

John, G. and Langley, P. (1995). Estimating continuous distributions in bayesian classifiers. In *In Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, pages 338–345. Morgan Kaufmann.

Lopes, G., Vidal, V., and Oliveira, M. (2016). 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*, pages 327–330. ACM.

Markos, Z., Doyore, F., Yifiru, M., and Haidar, J. (2014). 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.

Oliveira, M., Hairon, C., Andrade, O., Moura, R., Sicotte, C., Denis, J. L., Fernandes, S., Gensel, J., Bringel, J., and Martin, H. (2010). A context-aware framework for health care governance decision-making systems: A model based on the brazilian digital tv. In *2010 IEEE International Symposium on "A World of Wireless, Mobile and Multimedia Networks"(WoWMoM)*, pages 1–6.

ONU (2016). mortalidade infantil.

Piatetsky-Shapiro, G. (2005). Kdnuggets news on sigkdd service award 2005.

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

Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Ramos, R. F., Mattos, C. L. C., Júnior, A. H. S., Neto, A. R. R., Barreto, G. A., Mazzal, H. A., and Mota, M. O. (2016). 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-USA*.

Robu, R. and Holban, Ş. (2015). The analysis and classification of birth data. *Acta Polytechnica Hungarica*, 12(4).

ROSA, C. J. (2015). Aplicação de KDD nos dados dos sistemas SIM e SINASCf em busca de padrões descritivos de óbito infantil no município do rio de janeiro.