Revista Eletrônica de Sistemas de Informação
ISSN 1677-3071

v. 16, n. 2

mai-ago 2017 - Edição temática sobre Computação Urbana

DOI: https://doi.org/10.21529/RESI.2017.1602

Sumário

Computação Urbana

Revista hospedada em: www.periodicosibepes.org.br
Forma de avaliação: *double blind review*

Esta revista é (e sempre foi) eletrônica para ajudar a proteger o meio ambiente, mas, caso deseje imprimir esse artigo, saiba que ele foi editorado com uma fonte mais ecológica, a *Eco Sans*, que gasta menos tinta.

This journal is (and has always been) electronic in order to be more environmentally friendly. Now, it is desktop edited in a single column to be easier to read on the screen. However, if you wish to print this paper, be aware that it uses Eco Sans, a printing font that reduces the amount of required ink.

USANDO O CLASSIFICADOR NAIVE BAYES PARA GERAÇÃO DE ALERTAS DE RISCO DE ÓBITO INFANTIL

USING THE NAIVE BAYES CLASSIFIER TO GENERATE INFANT DEATH RISK ALERTS

(artigo submetido em agosto de 2017)

Cristiano Lima da Silva
Mestre pelo IFCE/UECE
Professor do IFCE-Tabuleiro do Norte
cristianocagece@gmail.com

Joyce Quintino Alves
Tecnico pelo IFCE
Bolsista FUNCAP no IFCE-Aracati
joycequintino11@gmail.com

Oton Crispim Braga
Graduado pelo IFCE
Bolsista FUNCAP/BPI no IFCE-Aracati
otoncbraga@gmail.com

José Wellington Pereira Júnior
Tecnico pelo IFCE
Bolsista FUNCAP/BPI no IFCE-Aracati
wj070996@gmail.com

Luiz Odorico Monteiro de Andrade
Doutorado pela UNICAMP
Professor da UFC- Sobral
odorico0811@gmail.com

Antônio Mauro Barbosa de Oliveira
Doutorado pela Univ. Pierre et Marie Curie
Professor do IFCE-Aracati
amauroboliveira@gmail.com

RESUMO

GISSA é um sistema inteligente para a tomada de decisões em saúde focado no cuidado materno infantil. Neste sistema, vários alertas são gerados nos cinco domínios da saúde (clínico-epidemiológico, normativo, administrativo, gestão do conhecimento, conhecimento compartilhado). O sistema se propõe a contribuir para a redução da mortalidade infantil no Brasil. Este artigo apresenta o LAIS, um mecanismo inteligente que usa aprendizado de máquina para gerar alertas de risco de mortalidade infantil no GISSA. Para tanto, este trabalho usa uma metodologia baseada na mineração de dados para alcançar um modelo de aprendizagem capaz de calcular a probabilidade de um recém-nascido morrer. Os testes mostram que o classificador Naive Bayes é o mais adequado para este propósito, apresentando bons resultados, com área da curva ROC de 92,1%. O trabalho reúne bases de dados do Ministério da Saúde, SIM e SINASC, para o treinamento de algoritmos de classificação, identificando relações entre dados de nascimento e de morte de crianças com menos de um an. Durante o processo metodológico foi utilizado o algoritmo *spread subsample*, que aplica sub-amostragem, melhorando os resultados do modelo.

Palavras-chave: Sistema de apoio à tomada de decisão; mineração de dados; Naïve Bayes; mortalidade infantil.

ABSTRACT

GISSA is an intelligent system for health decision making focused on children maternal care. Alerts are generated in this system that involve the five health domains: clinical-epidemiological, normative, administrative, knowledge management and shared knowledge. The system intends to contribute to the reduction of child mortality in Brazil. This paper presents LAIS, an intelligent mechanism that uses machine learning to generate child death risk alerts in GISSA. A data mining methodology is used to obtain a learning model able to calculate the probability of a newborn dying. The tests show that the Naive Bayes classifier is the most suitable algorithm for this purpose, presenting good results with a ROC curve of 92.1%. The work brings together the SIM and SINASC public databases for the training of classification algorithms, identifying relationships between birth and death data of children under one year of age. The spread subsample balancer was used, during the methodological process, which applies subsampling, improving model results.

Key-words: Decision support systems; data mining; Naïve Bayes; infant mortality.

1. INTRODUÇÃO

A mortalidade infantil é um problema que atinge principalmente os chamados países subdesenvolvidos (ONU, 2014). Segundo a organização das Nações Unidas (ONU), a taxa global de mortalidade infantil caiu 53% em 25 anos (ONU, 2015a), enquanto no Brasil esta redução foi de 77% nos últimos 22 anos (ONU, 2015b), provavelmente devido às melhorias no atendimento materno e infantil, por meio de programas de apoio a gestantes, como a Rede Cegonha, cujo objetivo é preservar a saúde da mãe e da criança, em especial nos primeiros anos de vida (VACONCELOS, 2013). O estado do Ceará teve uma redução de mortalidade infantil de 11,5% entre 2014 e 2015 (G. DO ESTADO DO CEARÁ, 2016). No entanto, os índices ainda são altos comparados com os de países desenvolvidos. A Noruega, por exemplo, apresentou no ano de 2014 uma taxa de mortalidade infantil de 2,4% (FACTBOOK, 2015). Portanto, estratégias mais eficazes são necessárias para amenizar esse problema.

Sistemas Inteligentes têm se tornado ferramentas cada vez mais importantes para auxiliar gestores no processo de tomada de decisão. Técnicas baseadas em aprendizado de máquina são capazes de prever alguns eventos com certa precisão (GOLDSCHMIDT *et al.*, 2015). A mineração de dados identifica padrões de comportamentos, gerando alertas para os gestores.

O GISSA (Governança Inteligente em Sistemas de Saúde) é uma plataforma inteligente de governança para o apoio à tomada de decisão em ambientes de saúde. O sistema é capaz de gerar alertas e relatórios administrativos para gestores e profissionais de saúde.

Este trabalho apresenta o LAIS, um mecanismo baseado em aprendizado de máquina capaz de prever casos de mortalidade infantil para auxiliar os gestores na tomada de decisão. É feita a integração e análise das bases públicas do SIM (Sistema de Informação sobre Mortalidade) e SINASC (Sistema de Informação sobre Nascidos Vivos), disponibilizadas pelo DATASUS (Departamento de Informática do Ministério da Saúde). O modelo gerado pelo LAIS é capaz de, a partir de atributos do recém-nascido e de sua mãe, calcular o risco de óbito infantil.

Este trabalho está organizado da seguinte forma. Na seção 1 são apresentados os projetos LARIISA e GISSA; na seção 2 são abordados os trabalhos relacionados ao LAÍS; na seção 3 é descrito o módulo inteligente baseado em aprendizado de máquina utilizando a metodologia de reconhecimento de padrões; e na seção 4 são apresentadas as conclusões e mencionados os possíveis trabalhos futuros.

2. FUNDAMENTAÇÃO TEÓRICA

2.1. O projeto LARIISA

LARIISA é uma plataforma desenvolvida em 2009 (OLIVEIRA *et al.*, 2010) com o objetivo de prover inteligência de governança nos cinco

domínios da saúde (clínico-epidemiológico, normativo, administrativo, gestão do conhecimento e conhecimento compartilhado). Ela ajuda os diversos usuários (pacientes, agentes de saúde, enfermeiros, médicos, pessoal administrativo, secretário de saúde, etc.) na tomada de decisões. Para tanto, faz-se necessário o manejo de bases de dados relacionadas à saúde, dispersas em bases governamentais ou não, cruzando-as com informações capturadas em tempo real (GARDINI *et al.*, 2013).

2.2. GISSA

O projeto GISSA é uma instância da plataforma LARIISSA, com foco no projeto Rede Cegonha do Ministério da Saúde do Brasil, suportado pela FINEP (Financiadora de Estudos e Projetos), em execução pelo Instituto Atlântico. Seu objetivo é apoiar tomadores de decisão, em todos os níveis do ciclo de saúde (pacientes, agentes de saúde, médicos, gestores de hospital, secretário, etc.), mediante a geração de alertas e *dashboards*, a partir da análise de dados nas diversas bases de saúde disponíveis, relacionadas à questão materno-infantil. Um protótipo do GISSA encontra-se operacional na cidade de Tauá, no Ceará e está sendo implantado em outros municípios do Estado.

O GISSA é formado por um conjunto de componentes que permite a coleta, integração, e visualização de informações relevantes ao processo de tomada de decisões (ANDRADE, 2015). Atualmente, ele dispõe dos seguintes alertas: nascido vivo com baixo peso; vacinação atrasada; relacionados ao pré-natal; campanha de vacina; entre outros. Nesse contexto, Freitas (2017) propõe um mecanismo baseado em heurísticas capaz de calcular a probabilidade de óbito de recém-nascidos usando informações de bases de dados distintas para o GISSA. No entanto, apesar de basear-se no conhecimento médico, o trabalho não realiza testes de eficiência ou precisão.

3. TRABALHOS RELACIONADOS

Considera-se que a desnutrição é um dos principais causadores de mortalidade infantil em países subdesenvolvidos. Em Markos (2014) foram utilizados algoritmos de classificação para encontrar padrões relativos ao estado nutricional de crianças menores de cinco anos. O estudo teve o objetivo identificar quais fatores afetam o estado nutricional das crianças. Foram tratados 11.654 casos com 16 atributos de saúde e socioeconômicos, coletados de uma pesquisa demográfica de saúde da Etiópia, realizada em 2011. Os algoritmos de aprendizado de máquina utilizados foram J48 (QUINLAN, 2014), Naive Bayes (JOHN, 1995) e o classificador de indução de regras PART (FRANK, 1998). Após diversos experimentos, foi selecionado o algoritmo PART, que apresentou o melhor desempenho, tendo precisão de 92,6% e área da curva ROC (*Receiver Operating Characteristic*) de 97,8%.

Em Rosa (2015) foi realizado um estudo sobre óbito infantil em crianças menores de um ano utilizando técnicas de mineração de dados. O

trabalho fez uso das bases de dados SIM e SINASC referentes ao Município do Rio de Janeiro entre os anos de 2008 e 2012. Realizou-se a integração por meio do campo DN (Número da Declaração de Nascimento), presente no SINASC e no SIM. Foram relacionados 3.336 indivíduos que nasceram e sofreram óbito infantil. Na pesquisa, foram usados os seguintes 13 atributos: sexo do RN (recém-nascido), Apgar1 e Apgar5, que são parâmetros de frequência cardíaca, respiração, tônus musculares, irritabilidade e cor da pele (avaliados durante o primeiro e quinto minuto de vida da criança, respectivamente), peso, cor do RN, idade do RN, causa básica da morte, idade da mãe, quantidade de filhos mortos, quantidade de filhos vivos, número de semanas de gestação, tipo da gravidez e tipo do parto. Utilizou-se o algoritmo não supervisionado Apriori (AGRAWAL *et al.*, 1994) para a investigação das características de nascimento associadas ao óbito em menores de um ano de idade. Ao final do trabalho, foram identificadas algumas regras que podem auxiliar os profissionais de saúde.

Em Robu (2015) foi apresentado um estudo sobre os nascimentos ocorridos no *Bega Obstetrics and Gynecology Clinique*, na Romênia, em 2010. Foi analisado um conjunto de dados com 2.325 nascimentos e 15 atributos: idade da mãe, número de gestações, número de semanas de gestação, sexo da criança, peso da criança e tipo de parto. O objetivo do trabalho era prever a pontuação do Apgar da criança ao nascer. Foram utilizados a ferramenta WEKA (WITTEN, 2016) e dez algoritmos de classificação: Naive Bayes, J48, IBK (WITTEN, 1991), Random Forest (AHA, 1991), SMO (PLATT, 1999), AdaBoost (FREUND *et al.*, 1996), LogitBoost (FRIEDMAN *et al.*, 2000), JRipp (COHEN, 1995), REPTree e SimpleCart (BREIMAN, 1993). O algoritmo LogitBoost apresentou melhores resultados nos experimentos. O modelo gerado foi usado em uma aplicação Java para prever a pontuação Apgar de um novo paciente.

Já Moreira (2016a) usou redes bayesianas para dar suporte à tomada de decisão em ambientes de incerteza. Uma rede foi desenvolvida para classificar desordens hipertensivas focada no cuidado da pré-eclâmpsia. Usando o modelo bayesiano Nisy-OR em uma base de dados, este sistema analisa a disposição dos dados e os classifica na rede. A partir dos sintomas apresentados pela gestante, o sistema, por meio de dados estatísticos, infere a gravidade do caso, ajudando o médico especialista no diagnóstico da pré-eclâmpsia. Esta abordagem mostrou-se precisa mesmo com um número pequeno de dados.

Moreira (2016b) fez uma análise detalhada entre os classificadores Naive Bayes e a árvore de decisão J48. O trabalho analisou um conjunto de dados relacionado a distúrbios hipertensivos, a fim de avaliar complicações na gravidez. Foi feito um estudo do desempenho dos classificadores e da matriz de confusão, usando parâmetros preditivos. Os dois classificadores apresentaram valores próximos. Contudo, a árvore de decisão J48 teve um resultado mais preciso.

4. MÓDULO INTELIGENTE BASEADO EM APRENDIZADO DE MÁQUINA

A fim de atingir melhores resultados no processo de mineração de dados, utilizou-se a metodologia de reconhecimento de padrões desenvolvida no Laboratório Centauro, da Universidade Federal do Ceará (UFC). Essa metodologia consiste em um conjunto de etapas do processo de mineração de dados, com o objetivo de selecionar os melhores algoritmos e atributos, de acordo com o contexto estudado (RAMOS, 2016).

A Figura 1 apresenta as etapas desenvolvidas neste trabalho: coleta dos dados; integração; avaliação e resultados; e aplicação. Primeiramente, os dados são coletados de bases de dados. Depois, esses dados são integrados por meio da junção entre as bases. Posteriormente, os algoritmos são treinados, testados e avaliados de acordo com a métrica apropriada, gerando um modelo de predição. Finalmente, o modelo gerado é testado em um protótipo capaz de predizer o risco de um recém-nascido vir a óbito.

4.1. Coleta e preparação de dados

Os dados foram coletados de duas bases públicas distintas: SINASC, que contém informações sobre nascidos vivos; e SIM, que contém informações sobre mortalidade, incluindo casos de mortalidade infantil. Ambas as bases estão disponíveis no portal do DATASUS no formato DBC *(DataBase Container*). Os dados são referentes ao estado do Ceará nos anos de 2013 (SINASC e SIM) e 2014 (SIM). Esses dados foram convertidos para SQL (*Structured Query Language*) usando o TABWIN, um software disponibilizado pelo DATASUS para visualização e manipulação de dados públicos.

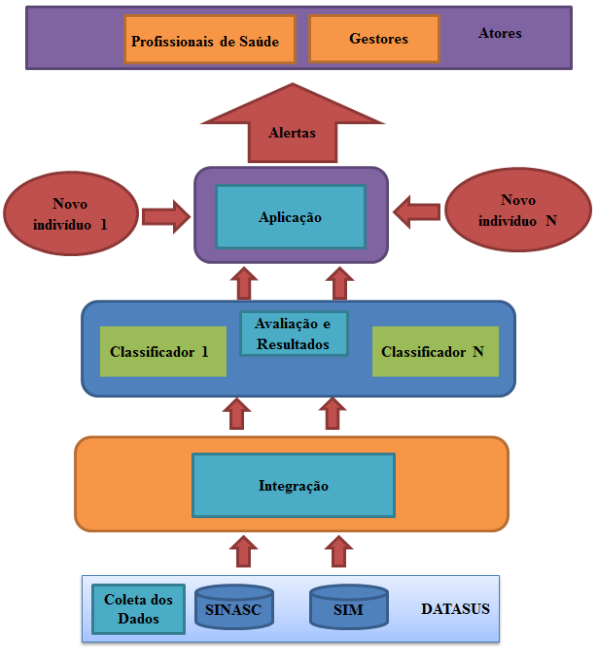

Figura 1. Metodologia para geração de alertas inteligentes.

Fonte: elaborada pelos autores.

4.2. Integração e seleção de atributos

Com o relacionamento das bases SIM e SINASC é possível recuperar informações sobre o nascimento de crianças vítimas de mortalidade infantil. Assim, é possível distinguir as crianças que sobreviveram, ou não, até atingirem um ano de idade.

Cada nascido vivo possui um atributo único chamado Número da Declaração de Nascido Vivo (numerodn), sempre preenchido na base do SINASC. A base do SIM também possui o campo (numerodn), que é preenchido somente em casos de óbito infantil. Esse campo foi essencial para a integração das bases, pois, a partir dele, pode-se relacionar os dados de mortalidade infantil aos dados de nascimento. A integração foi dividida em quatro etapas:

Etapa 1: levando em consideração que crianças nascidas em 2013 podem ser vítimas de mortalidade infantil em 2014, uniram-se as bases do SIM2013 e SIM2014, para crianças falecidas com menos de 1 an. A seguir, uma expressão simplificada (em álgebra relacional) do processo de integração (Equação 1):

$$SIM' \leftarrow \sigma idade \leq 1((SIM2013) \cup (SIM2014)) \quad (1)$$

Etapa 2: em seguida, foi feita a junção entre SINASC2013 e SIM' por meio do campo numerodn. O resultado retornou todos os casos de mortalidade infantil ocorridos em 2013 ou 2014. A seguir, é apresentada uma expressão simplificada (em álgebra relacional) (Equação 2):

$$M \leftarrow (\rho SN(SINASC) \rtimes SN.numerodn=S.numerodn\ \rho S(SIM')) \quad (2)$$

Etapa 3: também buscou-se casos de recém-nascidos que não sofreram óbito. Realizou-se uma consulta no SINASC2013, com exceção dos casos que sofreram óbito infantil (M). A seguir, é apresentada uma expressão simplificada em álgebra relacional (Equação 3):

$$V \leftarrow (SINASC2013 - (M)) \quad (3)$$

Etapa 4: por fim, fez-se uma união dos casos de óbito (M) e não óbitos (V) infantis ocorridos em 2013 e 2014. A seguir, é apresentada uma expressão simplificada em álgebra relacional (Equação 4):

$$TODAS \leftarrow (M \cup V) \quad (4)$$

Nessa etapa também foi realizada a rotulação dos casos, óbito (SIM) e não óbito (NÃO), necessária em problemas de classificação supervisionada.

Como resultado da integração dos dados, obteve-se um *dataset* com cinquenta atributos, contendo informações do nascimento e óbito (caso tenha ocorrido) de crianças nascidas em 2013. De acordo com o contexto, dezesseis atributos relacionados à mortalidade infantil foram selecionados, desprezando-se informações de características individuais, como endereço, identificadores etc. A partir dessa seleção, o novo *dataset* obtido resultou em 1.182 casos de óbitos e 124.876 casos de crianças que sobreviveram até um an. A Tabela I mostra os atributos selecionados para a etapa de análise e testes.

Atributos	Descrição
Idade	Idade da mãe
Estado civil	Estado civil da mãe
Escolaridade	Escolaridade da mãe
Local de nascimento	Local de ocorrência do nascimento
Filhos vivos	Quantidade de filhos vivos em gestações anteriores
Filhos mortos	Quantidade de filhos mortos em gestações anteriores
Tempo de gestação	Quantidade de semanas de gestação
Gravidez	Tipo de gravidez
Parto	Tipo de parto
Sexo	Sexo da criança
Peso	Peso ao nascer
Consultas	Quantidade de consultas no pré-natal.
Apgar1	Apgar 1 no primeiro minuto de vida
Apgar5	Apgar 5 no quinto minuto de vida
Anomalia	Com ou sem anomalia
Cor	Cor da criança

Quadro 1. Atributos selecionados

Fonte: elaborado pelos autores

4.3. Análise e testes

Com o intuito de se encontrar um modelo mais adequado para a predição de óbito infantil, foram realizados experimentos com os algoritmos de classificação *Random Forest* (RF), K *Nearest Neighbor* (KNN), Naive Bayes (NB), máquina de vetor de suporte (SVM), redes neurais artificiais (RNA) e J48, uma vez que esses se destacam na literatura com bons resultados. Estes algoritmos são descritos, a seguir:

(i) *K Nearest Neighbor* (KNN): calcula a similaridade entre o registro a ser analisado e os registros do conjunto de dados, a fim de estimar a classe do novo registro. Quando um novo registro deve ser classificado, ele é comparado a todos os registros do conjunto de treinamento para identificar k-vizinhos mais próximos, de acordo com alguma métrica selecionada, sendo que uma das mais utilizadas é a distância Euclidiana (Equação 5) e de Manhattan (Equação 6). Ao final, é selecionada a classe mais frequente entre os vizinhos mais próximos.

A distância Euclidiana refere-se à distância entre dois pontos medida pela linha reta que interliga esses dois pontos (Equação 5).

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

A distância Manhattan refere-se à distância entre dois pontos ao longo dos eixos coordenados, em ângulos retos (Equação 6).

$$d(x,y) = \sum_{i=1}^{n}|(x_{i-}y_i)| \quad (6)$$

(ii) Rede neural artificial: a variável x_i na equação 7 representa os sinais de entrada (dados sobre o problema), enquanto os pesos sinápticos são representados por w_i (responsável por ponderar os sinais de entrada de acordo com o nível de importância), e $\sum$ representa a função agregadora. Tem-se também +Θ, limiar de ativação, uma constante responsável por permitir ou não a passagem de sinal; u representa o potencial de ativação (SILVA, 2010).

$$u = \sum_{j=1}^{n} w_{i*}x_i + \Theta \quad (7)$$

(iii) Máquina de vetor de suporte - SVM: tem origem na aplicação do aprendizado estatístico (CORTES, 1995). Os algoritmos SVMs constroem os denominados classificadores lineares, que separam o conjunto de dados por meio de um hiperplano; uma reta divide o conjunto de dados em dois subconjuntos. Durante o processo de mineração de dados são geradas várias retas que dividem o conjunto de dados. O algoritmo SVM deve realizar o processo de escolha da reta separadora. Dessa forma, é criado um vetor otimizado que é, então, utilizado para classificar novas instâncias.

(iv) Naive Bayes: o algoritmo Naive Bayes é baseado na teoria das probabilidades e supõe que os atributos vão influenciar a classe de modo independente. Durante a criação do modelo, o classificador constroi uma tabela mostrando quanto cada categoria de cada atributo contribui para cada classe. Na equação 8, C representa a classe $\{A_1 = a_1, \ldots \{A_n = a_n\}$.

$$P(A_1, \ldots A_n, C = P(C) \prod_{k=1}^{n} P(A_i|C) \quad (8)$$

(v) J48: esse algoritmo gera uma estrutura de decisão em formato de árvore que, a cada nova instância, é percorrida até se chegar ao nó folha onde está a classe. Da mesma forma, a árvore começa com um único nó raiz e depois vai sendo dividida até chegar à classe. Da mesma forma, a árvore começa com um único nó raiz e depois vai sendo dividida até chegar à classe.

Esse algoritmo é constituído de duas fases: construção e simplificação. Durante a fase de construção ocorre um particionamento recursivo do conjunto de dados de treinamento. Há duas operações que são realizadas durante esse processo: avaliação dos pontos de separação em potencial e criação das partições. Após definir o melhor ponto de separação de cada nó, podem ser criadas partições pela aplicação do critério de separação identificado.

(vi) *Random Forest*: o algoritmo *Random Forest* cria várias árvores de decisão que depois são utilizadas na classificação de novos objetos. Cada conjunto de árvores passa por um mecanismo de votação, que elege a classificação mais votada. A classificação encontra-se nos nós terminais das árvores, tornando, dessa maneira, esse algoritmo mais poderoso do que uma árvore de decisão simples.

4. AVALIAÇÃO E RESULTADOS

Com o objetivo de identificar os melhores algoritmos para o contexto de predição de mortalidade infantil, realizou-se o treinamento e testes dos algoritmos destacados na literatura. Para tanto, foi adotado o método de validação cruzada, dividindo a base de dados em dez partes. O método é baseado em estratificação, o que reduz a variância estimada, além de evitar altos custos computacionais (JAPKOWICZ, 2011). A Tabela 1 mostra os resultados obtidos no experimento I com o uso de alguns classificadores. Para analisar os resultados, utilizou-se a área da curva ROC como métrica de avaliação. Trata-se de uma das métricas mais utilizadas para a avaliação de classificadores. Ela leva em consideração tanto os casos verdadeiros positivos (paciente que sofreu óbito e foi classificado como óbito) quanto os falsos positivos (paciente vivo que foi classificado como óbito).

Tabela 1. Experimento I

Algoritmos	*Precision*	*Recall*	*F-Measure*	*ROC*
J48	0,671	0,292	0,409	0,808
R. FOREST	0,640	0,289	0,399	0,883
NAIVE BAYES	0,294	0,607	0,396	0,921
IBK	0,479	0,273	0,348	0,785
V. PERCEPTRON	0,695	0,285	0,404	0,642
MLP	0,689	0,287	0,405	0,911
SMO	0,567	0,306	0,398	0,857

Fonte: elaborada pelos autores

O quantitativo de crianças que sobrevivem é bastante superior ao quantitativo de crianças que morrem antes de completar um ano de idade. Por isso, há um problema conhecido como desequilíbrio de classes, causando mal desempenho dos classificadores. Neste caso, as chances de ocorrência de *overfitting* e possíveis classificações incorretas de novos casos de óbitos (classe “SIM”) são altas. Para diminuir essas chances, fez-se outro experimento (tabela III) usando o algoritmo de balanceamento de dados *Spread Subsample* para balancear as classes por meio de subamostragem aleatória. Após o balanceamento, o número de indivíduos das classes “SIM” e “NAO” foi 1.182 instâncias para cada classe. Com o uso do balanceamento de dados, alcançou-se resultados mais precisos.

Tabela 2. Experimento II

Algoritmos	*Precision*	*Recall*	*F-Measure*	*ROC*
J48	0,895	0,830	0,861	0,888
R. FOREST	0,873	0,831	0,851	0,913
NAIVE BAYES	0,921	0,809	0,861	0,924
IBK	0,827	0,744	0,783	0,843
V. PERCEPTRON	0,900	0,838	0,868	0,875
MLP	0,837	0,821	0,829	0,898
SMO	0,882	0,843	0,826	0,865

Fonte: elaborada pelos autores

A tabela 3 mostra a matriz de confusão do Naive Bayes, o algoritmo que apresentou melhores resultados. Verifica-se que o Naive Bayes classificou corretamente 2.056 crianças (86,912%) que correspondem à diagonal de acerto na tabela abaixo (956 + 1.100). Portanto, 308 crianças (13,028%) foram classificadas incorretamente (outra diagonal: 82 + 226).

Tabela 3. Matriz de confusão do algoritmo Naive Bayes

		Classe predita	
		Morto	Vivo
Classe Real	Morto	956	226
	Vivo	82	1.100

Fonte: elaborada pelos autores

Entre as 308 crianças que foram classificadas erroneamente, 82 (3,46%) são falsos positivos e 226 (9,56 %) são falsos negativos. Das 2.056 crianças que foram classificadas corretamente, 956 (40,44 %) são verdadeiros positivos e 1.100 (46,53 %) são verdadeiros negativos.

Apesar dos erros na classificação, o Naive Bayes teve um desempenho satisfatório. Esse desempenho deve-se ao fato de ele considerar cada atributo de forma independente, conseguindo trabalhar com informações incompletas ou imprecisas (FACELI, 2011), diminuindo as chances de ocorrência de *overfitting* e má classificação dos casos de óbitos. Porém, em casos em que existem relacionamentos entre os atributos, fato muito comum em problemas de saúde, o resultado acaba sendo prejudicado (MOREIRA, 2017). Assim, a probabilidade de um recém-nascido ir a óbito dado um espaço amostral de recém-nascidos foi estimada com maior precisão.

O Naive Bayes teve um percentual de acerto de 86%. Para melhorar a classificação de dados e, consequentemente, aumentar a taxa de acerto, como trabalho futuro, outros atributos (informações sobre a mãe possuir doença cardíaca ou ser hipertensa, diabética, fumante, ou usuária de drogas ou álcool) pertencentes ao e-SUS poderão ser incluídos no *dataset*.

Após um processo de análise e comparação entre os algoritmos, fez-se a escolha do algoritmo de classificação mais eficiente de acordo com o domínio estudado. O classificador Naive Bayes foi o que melhor se adaptou ao conjunto de dados analisado. Em seguida, utilizou-se o modelo gerado pelo algoritmo para classificar o risco de novos pacientes sofrerem óbito e, a partir disso, calcular a probabilidade de isso ocorrer.

O cálculo dessa probabilidade foi dado pelo uso de um método pertencente à classe Naive Bayes existente na API disponibilizada pelo WEKA. Esse método, chamado de *distributionForInstance*, retorna a probabilidade de cada classe, ou seja, após o algoritmo classificar a nova instância como

“SIM” ou “NAO” essa instância é passada como parâmetro para o método que retorna a probabilidade de a classificação estar correta.

5. CONCLUSÃO E TRABALHOS FUTUROS

O LAIS, apresentado neste trabalho, agrega valor aos alertas do GISSA, dotando-os de um mecanismo inteligente baseado em classificadores. Assim, ele é capaz de fornecer ao gestor de saúde, importantes alertas sobre a probabilidade de óbito de um recém-nascido, a partir das informações da gestante e do próprio recém-nascido. Portanto, o tomador de decisão pode 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 a partir da integração das bases de dados SINASC e e-SUS realizada por Lopes (2016) e comparar os resultados com aqueles obtidos em Freitas (2017). A expectativa é que ambas as metodologias, classificação e heurísticas se adequem a determinadas classes específicas de problemas. Finalmente, há também a possibilidade do desenvolvimento de um mecanismo híbrido a ser agregado ao GISSA, a partir dessas duas experiências.

AGRADECIMENTOS

Os autores agradecem à PRPI/IFCE e à FUNCAP o apoio recebido via o programa Bolsa de Produtividade em Pesquisa, Estímulo à Interiorização e à Inovação Tecnológica.

REFERÊNCIAS

AGRAWAL, R.; SRIKANT, R. *et al.* “Fast algorithms for mining association rules,” in *Proc. 20th Int. Conf. Very Large Data Bases (VLDB)*, September 12-15, Santiago, Chile, v. 1215, p. 487–499, 1994.

ANDRADE, L. O. M.; OLIVEIRA, M.; RAMOS, R. “Projeto GISSA: META FÍSICA 3 – atividade 3.1 Definir modelo de inteligência de gestão na saúde”, Disponível em: https://amauroboliveira.files.wordpress.com/2015/11/2015-nov30-meta-3-ativ-1-moldelointeligc3aanciagestc3a3o-draf-1-0.pdf, 2015, Acesso em 30-September-2016.

BREIMAN, L.; FRIEDMAN, J.; OLSHEN, R.; STONE, C., “Classification and regression trees, Wadsworth International Group, Belmont, CA, 1984,” Case Description Feature Subset Correct Missed FA Misclass, v. 1, p. 1–3, 1993.

COHEN, W. W.; “Fast effective rule induction” In *Proceedings Of The Twelfth International Conference on Machine Learning*, July 9—12, Tahoe City, CA, USA, p. 115–123, 1995.

CORTES, C.; VAPNIK, V. “Support-vector networks,” Machine Learning, v. 20, n. 3, p. 273–297, 1995.

FACELI, K.; LORENA, A. C.; GAMA, J.; CARVALHO, A. “Inteligência Artificial: Uma abordagem de aprendizado de máquina”, LTC, p. 70, 2011.

FACTBOOK, C. W. “Noruega taxa de mortalidade infantil”. 2015. Disponível em: http: //www.indexmundi.com/pt/noruega/taxa de mortalidade infantil.html/. Acesso em 19 de Julho de 2017.

FRANK, E.; WITTEN, I. H. “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., p. 144–151, 1998.

FREITAS, R.; LIMA, C.; BRAGA, O.; LOPES, G.; OLIVEIRA, M.; ANDRADE O.; “Using linked data in the integration of data for the 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 17–18, Gramado, RS, Brazil. ACM, p. 327–330, 2017.

FREUND, Y.; SCHAPIRE, R. E. *et al.* “Experiments with a new boosting algorithm” in *ICML*, v. 96, p. 148–156, 1996.

FRIEDMAN, J.; HASTIE, T.; HASTIE, R., *et al.*, “Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors)”. *The Annals of Statistics,* v. 28, n. 2, p. 337–407, 2000.

G. DO ESTADO DO CEARÁ.“Ceará reduz mortalidades materna, infantil e fetal”. 2016. Disponível em: http://www.saude.ce.gov.br/index.php/noticias/47723-ceara-reduz-mortalidades-materna-infantil-e-fetal/, Acesso em 30 de Junho de 2017.

GARDINI, L. M.; BRAGA, R.; BRINGEL, J.; OLIVEIRA, C.; ANDRADE, R.; MARTIN, H.; ANDRADE, L. O.; OLIVEIRA, M.; “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, p. 334–339, 2013.

GOLDSCHMIDT, R.; PASSOS, E.; BEZERRA, E. *Data mining: um guia prático, conceitos, técnicas, ferramentas, orientações e aplicações*, 2. ed. Elsevier, 2015.

JAPKOWICZ, N.; SHAH, M. *Evaluating learning algorithms: a classification perspective.* Cambridge University Press, 2011.

JOHN, G. H; LANGLEY, P. “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., p. 338–345, 1995.

LOPES, G.; VIDAL, V.; OLIVEIRA, M., “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)*, Teresina, PI, Brazil. ACM, p. 327–330, Nov., 2016.

MARKOS, Z.; DOYORE, F.; YIFIRU, M.; HAIDAR, J.; “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*, v. 5, p. 152, 2014.

MOREIRA, M. W. L.; RODRIGUES, J. J. P. C.; OLIVEIRA, A. M. B.; RAMOS, R. F.; SALEEM, K. “A preeclampsia diagnosis approach using Bayesian networks” in *2016 IEEE International Conference on Communications (ICC)*, p. 1–5, May 2016.

MOREIRA, M. W. L.; RODRIGUES, J. J. P. C.; OLIVEIRA, A. M. B.; SALEEM, K.; NETO, A., “Performance evaluation of predictive classifiers for pregnancy care” *in 2016 IEEE Global Communications Conference* (GLOBECOM), p. 1-6, Dec 2016.

MOREIRA, M. W. L.; RODRIGUES, J.; BRINGEL, J.; OLIVEIRA, M.; SALEEM, K.; NETO, A. J. V. “Performance Evaluation of the Tree Augmented Naïve Bayes Classifier for Knowledge Discovery in Healthcare Databases” in *XVII Workshop de Informática Médica (17º WIM) – CSBC,* July 03 – 05, São Paulo, Brasil, 2017.

OLIVEIRA, M. *et al.*, "A context-aware framework for health care governance decision-making systems: A model based on the Brazilian Digital TV". *2010. IEEE International Symposium on "A World of Wireless, Mobile and Multimedia Networks" (WoWMoM), Montreal, QC,* Canada, p. 1-6, 2010, doi: 10.1109/WOWMOM.2010.5534979

ONU. “ONU afirma que taxa de mortalidade infantil no mundo caiu pela metade em 25 anos”. 2015. Disponível em: http://www.uai.com.br/app/noticia/saude/2015/09/09/noticias-saude,187094/onu-afirma-que-taxa-de-mortalidade-infantil-no-mundo-caiu-pela-metade. Acesso em 17 de Julho de 2017.

ONU. “ONU: Meta global de mortalidade infantil ’será atingida com atraso de 11 anos”. 2014. Disponível em: http://www.bbc.com/portuguese/noticias/2014/09/140916 unicef meta mortalidade infantil rm/. Acesso em 22 de Julho de 2017.

ONU. “Taxa de mortalidade infantil no Brasil cai 77% em 22 anos, diz ONU”. 2015. Disponível em: https://istoe.com.br/324257_TAXA+DE+MORTALIDADE+INFANTIL+NO+BRASIL+CAI+77+EM+22+ANOS+DIZ+ONU/. Acesso em 28 de Junho de 2017.

PLATT, J. C.; SCHÖLKOPF, B.; BURGES, C. J. C.; SMOLA, A. J. “Fast training of support vector machines using sequential minimal optimization,” Eds. Cambridge, MA, USA: MIT Press, p. 185–208. [Online]. Disponível em: http://dl.acm.org/citation.cfm?id=299094.299105, 1999

QUINLAN, J. R. *C4.5: programs for machine learning.* Elsevier, 2014.

RAMOS, R. F.; MATTOS, C. L. C.; JÚNIOR, A. H. S.; NETO, A. R. R.; BARRETO, G. A.; MAZZAL, H. A.; MOTA, M. O., “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.

ROBU, R.; HOLBAN, S. "The analysis and classification of birth data," in *Acta Polytechnica Hungarica*, v. 12, n. 4, 2015.

ROSA, C. J. "Aplicação de KDD nos dados dos sistemas SIM e SINASC em busca de padrões descritivos de óbito infantil no município do Rio de Janeiro,", Centro de Ciências Exatas e Tecnologia, Universidade Federal do Estado do Rio De Janeiro, Rio de Janeiro, 2015.

SILVA, I. D.; SPATTI, D. H.; FLAUZINO, R. A.,"*Redes neurais artificiais para engenharia e ciências aplicadas*," São Paulo: Artliber, p. 33–111, 2010.

VACONCELOS, A. L. R.; GUERRERO, A. V. P. "Rede cegonha". In: *Physis Revista de Saúde Coletiva*, v. 23, n. 4, p. 1297–1316, 2013.

WITTEN, D. W.; KIBLER, D.; ALBERT, M. K. "Instance-based learning algorithms". *Machine Learning*, v. 6, n. 1, p. 37–66, 1991. [Online]. Disponível em: http://dx.doi.org/10.1007/BF00153759

WITTEN, I. H.; FRANK, E.; HALL, M. A.; PAL, C. J. Data mining: practical machine learning tools and techniques. Morgan Kaufmann, 2016.