

Triagem automática de gestantes de alto risco de mortalidade usando árvores de decisão*

Juliane Pires Martins¹[0000-0003-4092-0656] e
Sandro da Silva Camargo¹[0000-0001-8871-3950]

Programa de Pós-Graduação em Computação Aplicada (PPGCAP)
Universidade Federal do Pampa (UNIPAMPA)
Av. Maria Anunciação Gomes do Godoy, 1650 - 96413-170, Bagé- RS - Brasil
<https://www.cursos.unipampa.edu.br/cursos/ppgcap/>
{julianemartins.aluno, sandrocamargo}@unipampa.edu.br

Resumo. O risco de mortalidade materna é um grande problema de saúde nos países em desenvolvimento. Para reduzir a mortalidade, podem ser desenvolvidas estratégias de intervenção mais abrangentes, com a aplicação de testes de triagem, visando reduzir complicações relacionadas à gravidez. Este trabalho visa construir um modelo preditivo baseado em dados de diferentes hospitais e clínicas comunitárias para fazer uma previsão de risco de mortalidade em mulheres grávidas através de seis preditores de risco. O modelo desenvolvido obteve uma acurácia de 71,78%, com uma sensibilidade, para predizer pacientes de alto risco de mortalidade, de 87,04%.

Palavras-Chave: Mineração de Dados · Classificação · Saúde Pública · Gestação.

1 Introdução

Segundo uma iniciativa global das Nações Unidas, dentro dos Objetivos do Desenvolvimento Sustentável (ODS), foi inserida uma meta para reduzir as taxas de mortalidade materna evitável de 2016 a 2030. No mundo, a cada dia, cerca de 830 mulheres morrem por causas evitáveis relacionadas ao parto e gravidez, e 99% de todas as mortes maternas ocorrem em países em desenvolvimento [1].

Embora os cuidados pré-natal e o acompanhamento médico antes e durante gravidez diminuam significativamente o risco de complicações, ainda há margem para redução dos fatores de risco, o que é essencial na fase inicial dos sintomas [2]. Dessa forma, analisar fatores como idade, pressão arterial, frequência cardíaca, níveis de glicose, dentre outros, pode aumentar o conhecimento sobre o nível de risco das mulheres na gravidez e, assim, auxiliar o planejamento de medidas preventivas.

Um dos ramos emergentes na área de sistemas de informação em saúde é o apoio à decisão clínica que auxilia os cuidados de saúde ao contribuir com

* O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.

J. P. Martins; S. S. Camargo

o processo de diagnóstico, fornecendo informações úteis, baseadas em conhecimento obtido a partir de dados históricos. Este tipo de sistema não pretende substituir ou tomar decisões pelos médicos, mas é desenvolvido com o objetivo de ajudar a melhorar a processo de decisão clínica[3]. Como consequência deste desenvolvimento, a bibliografia tem mostrado bons resultados através de técnicas de previsões por árvores de decisão. Um classificador baseado em árvores de decisão, criado a partir de um determinado conjunto de dados históricos, gera um modelo composto por conjunto de regras que permite prever a classe, ou diagnóstico, para novas amostras [4].

Neste contexto, o objetivo deste trabalho é construir um modelo de classificação, baseado em árvores de decisão, para prever o risco de morte em mulheres grávidas. Os dados foram coletados em diferentes hospitais, clínicas comunitárias, e centros de cuidado de saúde materna das áreas rurais de Bangladesh por meio de um sistema de monitoramento de risco baseado em Internet das Coisas (IoT) [2].

Este artigo está organizado da seguinte forma: Na Seção 2 são apresentados o Material e os Métodos. Na Seção 3 são abordados os Resultados e Discussões. O trabalho será finalizado na Seção 4, com as conclusões e propostas de trabalhos futuros.

2 Material e Métodos

2.1 Base de Dados

Os dados utilizados neste estudo foram coletados de diferentes hospitais, clínicas comunitárias, e centros de cuidado de saúde materna das áreas rurais de Bangladesh por meio de um estudo de sistema de monitoramento de risco baseado em IoT [5]. A coleta foi realizada em Dhaka, Bangladesh, no período compreendido entre janeiro de 2018 e junho de 2020. Foi utilizado um conjunto de dados já existente de um software de mineração de dados chamado Waikato Environment for Knowledge Analysis (Weka), desenvolvido pela 'University of Waikato' na Nova Zelândia. Duas funcionalidades do Weka foram aplicadas, a primeira para mineração de informações de conjuntos de dados médicos de pacientes com parâmetros de idade, nível de glicose, pressão arterial, frequência cardíaca, temperatura corporal, dentre outros; e outra para classificar o nível de risco com precisão com a ajuda do algoritmo de aprendizado de máquina a melhor previsão do nível de risco [2].

Três categorias de risco de mortalidade materna foram consideradas: baixo risco, médio risco e alto risco. Foram coletados dados de 1.014 pacientes, das quais 406 foram classificadas como de baixo risco, 336 classificadas como médio risco e 272 de alto risco [2].

Sete preditores foram coletados para cada paciente:

1. Age: Idade em anos das gestantes;
2. SystolicBP: Valor superior da pressão sanguínea, medida em mmHg;
3. DiastolicBP: Valor mais baixo da pressão sanguínea, medido em mmHg;

Triagem automática de gestantes de alto risco de mortalidade

4. BS: nível de glicose no sangue, medido em termos de concentração molar, mmol/L;
5. BodyTemp: temperatura corporal;
6. HeartRate: número de batimentos cardíacos por minuto, em situação de repouso; e
7. Risk Level: nível de risco de mortalidade materna diagnosticado durante a gestação, baseado nos preditores anteriores, que é a variável nominal para classe a ser predita.

2.2 Ferramentas

Para o desenvolvimento deste estudo, foi utilizada a linguagem R, através do programa R-Studio, versão 4.1.1, instalado em um sistema operacional Windows 10 Pro - 64 bits com um processador AMD Ryzen 5 5600X 16GB de memória RAM e disco sólido interno SSD 128GB. Também foi utilizado o pacote *Classification and Regression Training* (Caret) versão 6.0-86, *Rpart.plot* versão 3.1.1.

2.3 Árvores de Decisão

Neste trabalho, foram aplicadas as técnicas de aprendizado de máquina supervisionado, que são uma classe de algoritmos que visam aprender uma função desconhecida que mapeia os dados de entrada a um conjunto de classes de saída, previamente determinadas por especialistas do domínio. Após o processo de aprendizado supervisionado, é criado um modelo que pode ser utilizado para simular o conhecimento dos especialistas.

Árvores de decisão são um dos métodos de aprendizado supervisionado amplamente utilizados e práticos de classificação [6]. A motivação para sua utilização é que são modelos do tipo caixa branca, com fácil interpretabilidade [7]. Uma árvore de decisão é um fluxograma do tipo árvore, no qual cada nó interno denota um teste de um preditor, cada ramo representa um resultado de teste e cada nó folha (nó terminal) tem um nome de classe [8]. A posição dos preditores no modelo criado ocorre de acordo com a sua importância para predição da classe de saída. Quanto mais relevante o preditor, mais próximo ele estará do topo da árvore, ou seja, mais próximo da raiz e mais distante das folhas. O algoritmo de árvores de decisão utilizado foi o *J48* para determinar a classificação da precisão do nível de risco como alto, médio ou baixo.

3 Resultados e Discussão

Os resultados aqui apresentados são discutidos em duas perspectivas: uma análise estatística preliminar seguida pela criação dos modelos de árvores de decisão.

J. P. Martins; S. S. Camargo

3.1 Análise Estatística

A Tabela 1 apresenta a estatística descritiva dos dados utilizados neste estudo. Esta tabela apresenta, para cada preditor, o valor mínimo (Min), primeiro quartil (Q1), mediana, média, terceiro quartil (Q3), valor máximo (Max), desvio padrão (DP), coeficiente de variação (CV), p-valor do teste de normalidade de Shapiro Wilk (Norm) e o valor F do teste de médias de Tukey (Valor F) para diferenças estatisticamente significativas entre pacientes dos três grupos: baixo, médio e alto risco de mortalidade.

Tabela 1. Resumo das estatísticas descritivas

Preditores	Min	Q ₁	Mediana	Média	Q ₃	Max	DP	CV	Norm	Valor F
Age	10	19	26	29.872	39	70	13.474	0.451	0	46.114
SystolicBP	70	100	120	113.198	120	160	18.404	0.163	0	95.903
DiastolicBP	49	65	80	76.461	90	100	13.886	0.182	0	84.960
BS	6	6.9	7.5	8.726	8	19	3.294	0.377	0	331.160
BodyTemp	98	98	98	98.665	98	103	1.371	0.014	0	16.446
HeartRate	7	70	76	74.302	80	90	8.089	0.109	0	20.453

Os resultados mostram que todos os preditores apresentam distribuição normal dado que os valores P do teste de Shapiro-Wilk são menores que 0.05 ($p < 0.05$). Em consequência deste resultado, o teste paramétrico de médias de Tukey foi usado para identificar se os grupos tem diferenças estatisticamente significativas entre suas distribuições, com nível de significância de 5%, ou $p < 0.05$. Foram apontadas diferenças estatisticamente significativas para os três grupos em todos os preditores, com ênfase para os maiores valores F do Teste de Tukey em BS e SystolicBP. A Figura 1 apresenta as distribuições de densidade dos seis preditores em relação aos diferentes grupos. Verifica-se que os valores de BS, estão distribuídos similarmente ao longo do gráfico para alto risco, e tem valores bem concentrados em valores inferiores a 8 para as demais classes. Já os valores de SystolicBP tem uma distribuição mais concentrada em valores superiores a 135 para alto risco. Já o médio e baixo risco estão mais concentrados em valores inferiores a 135.

3.2 Árvores de Decisão

Para criar o modelo de árvore de decisão, primeiramente as 1.014 amostras foram divididas em dois conjuntos: dados de treino, com 80% das amostras, e dados de teste, com os demais 20% das amostras. Esta divisão foi feita de forma aleatória estratificada. O conjunto de treino continha 812 amostras, sendo 218 de alto risco, 269 de médio risco e 325 de baixo risco.

A Figura 2 mostra o modelo de árvore de decisão criado com base nos dados de treino. O modelo mostra que quando o índice de glicose da amostra for maior ou igual a 8, ela terá 76% de probabilidade de ser de alto risco, sendo que 26%

Triagem automática de gestantes de alto risco de mortalidade

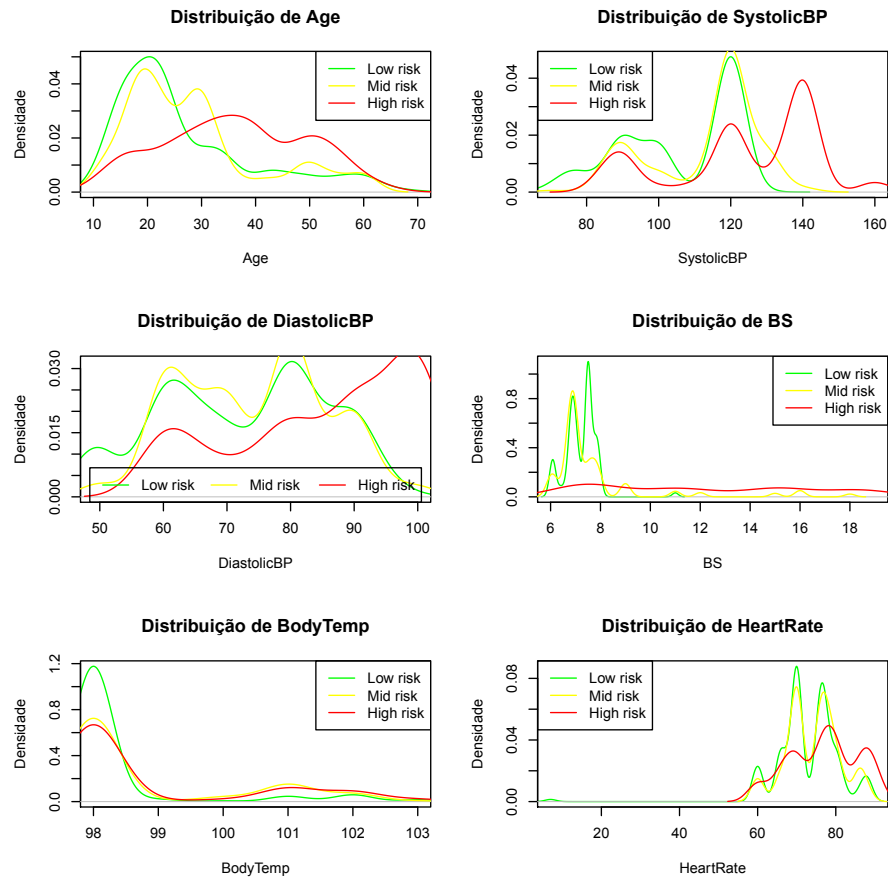
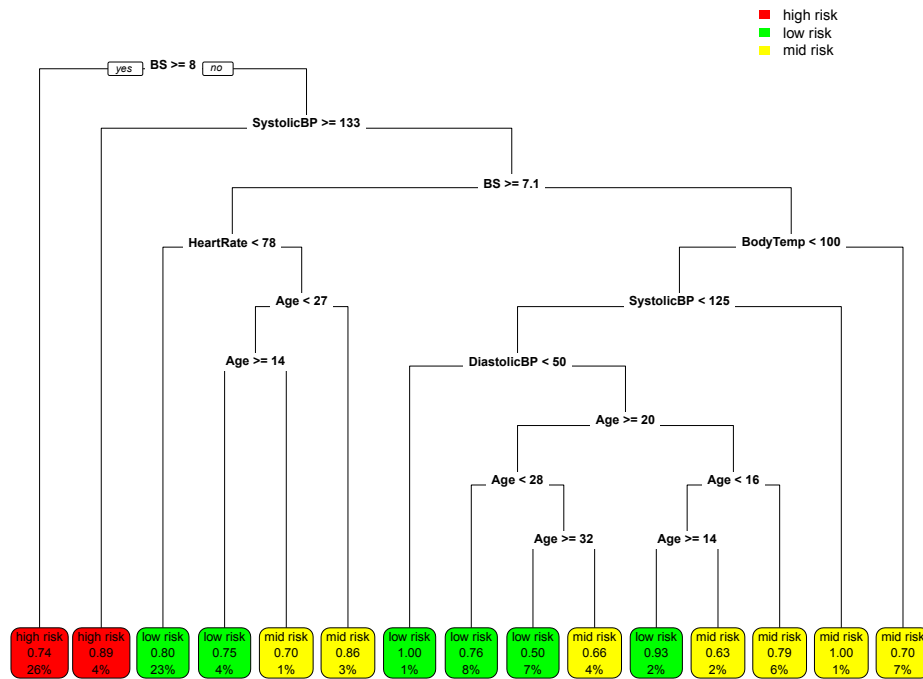


Fig. 1. Densidade de distribuição dos preditores por classe.

das amostras enquadraram-se na regra representada neste ramo da árvore. Caso o nível de glicose não seja maior ou igual a 8; e a pressão arterial sistólica seja maior ou igual a 133, a amostra terá 91% de probabilidade de ter alto risco, sendo que 4% das amostras enquadraram-se na regra representada neste ramo da árvore. Estas duas folhas da árvore já permitem classificar as amostras de alto risco, que são a maior preocupação em termos de triagem.

Mesmo que a maior preocupação seja a triagem das amostras de alto risco, também há outros ramos da árvore que expressam regras importantes para as outras classes. Por exemplo, o terceiro ramo da árvore expressa uma regra de que se o BS estiver no intervalo entre 7.1 e 8, se SystolicBP não for maior ou igual a 133, e se HeartRate é menor que 78, a amostra será classificada como de baixo risco. Esta regra tem 80% de acerto para os 23% de amostras de baixo risco que se enquadraram nela. Os demais ramos da árvore são menos relevantes porque um percentual menor ou igual a 8% das amostras está enquadrado neles.

J. P. Martins; S. S. Camargo

Fig. 2. Modelo de Árvore de Decisión criado pelo algoritmo *J48*.

Após o treinamento do modelo, sua capacidade preditiva foi testada sobre os dados de teste, que eram compostos de 202 amostras, sendo 54 de alto risco, 67 de médio risco e 81 de baixo risco. A Figura 3 apresenta o resultado. A área sombreada em vermelho mostra os ramos da árvore que se enquadram na predição de amostras de alto risco.

A matriz de confusão na Tabela 2 mostra as classificações realizadas nos dados de teste. Existem 54 amostras de risco alto, das quais 47 foram classificadas corretamente pelo modelo e 7 classificadas incorretamente. Das 67 amostras de risco médio, 26 foram classificadas de forma correta e 41 de forma incorreta. Por fim, com risco baixo, das 81 amostras existentes, 72 estão corretamente classificadas e 9 incorretamente classificadas. A matriz de confusão mostra que a maior dificuldade de predição é para as amostras de risco médio.

A Tabela 3 sumariza as métricas do modelo a partir dos dados de teste. Os resultados mostraram uma acurácia de 71,78%. Com um intervalo de confiança de 95%, estima-se que o modelo pode atingir uma acurácia entre 66,04% a 77,87%. A taxa sem informação de 40,1% indica a proporção de amostras da maior classe, que seria o *baseline* de desempenho, caso o modelo indicasse todas as amostras como a classe mais populosa. O coeficiente Kappa é de 0,5688, resultando em um índice aceitável das decisões do modelo. O Valor P do teste de McNemar indica um modelo estatisticamente significativo. Estas métricas podem ser aprimoradas com uma disponibilidade maior de amostras de treinamento.

Triagem automática de gestantes de alto risco de mortalidade

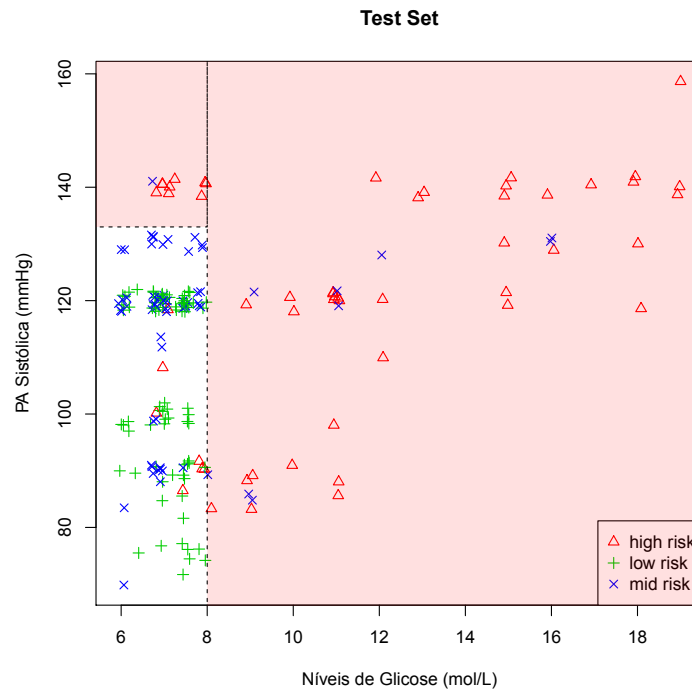


Fig. 3. Superfície de Decisão da árvore representada nas amostras de teste.

As demais métricas são analisadas individualmente por classe. Valores estão na faixa de valores entre 0 e 1, sendo que valores mais próximos ao limite máximo indicam um melhor resultado do modelo. A sensibilidade, que representa a capacidade do modelo de prever casos positivos, é de 0,8704 para risco alto, 0,3881 para risco médio e 0,8889 para risco baixo. A boa sensibilidade para amostras de risco alto é um aspecto altamente desejável para o modelo, já que amostras nesta classe tem maior possibilidade de mortalidade materna. A maior sensibilidade para o risco baixo pode ter relação com a maior quantidade de amostras de treinamento para esta classe, já que há 50% mais amostras de baixo risco em relação à quantidade de amostras de alto risco. O menor valor de sensibilidade para a classe de risco médio pode ser devido a não haver um padrão claro para reconhecer amostras desta classe.

A especificidade, que é a capacidade do modelo detectar casos negativos, indica 0,8919 para risco alto, 0,8963 para risco médio e 0,7769 para risco baixo. Estes valores indicam altas probabilidades do modelo prever amostras que não são de risco alto também são um valor desejável muito positivo para o modelo. O mesmo ocorre para amostras de risco médio. O valor preditivo positivo indica a probabilidade de pertencimento a uma classe quando o modelo prediz esta

J. P. Martins; S. S. Camargo

Tabela 2. Matriz de Confusão.

		Classes Reais			Totais
		Risco Alto	Risco Médio	Risco Baixo	
Classes Preditas	risco alto	47	16	0	63
	risco médio	5	26	9	40
	risco baixo	2	25	72	99
Totais		54	67	81	202

Tabela 3. Métricas do modelo.

Métrica	Valor
Acurácia	0,7178
95% CI	(0,6504, 0,7787)
Taxa sem informação	0,401
P-Value [Acc > NIR]	< 2.2e-16
Kappa	0,5688
P-Value Test Mcnemar's	0,001584

Métrica por Classe	risco alto	risco médio	risco baixo
Sensibilidade	0,8704	0,3881	0,8889
Especificidade	0,8919	0,8963	0,7769
Valor Preditivo Positivo	0,7460	0,6500	0,7273
Valor Preditivo Negativo	0,9496	0,7469	0,9126
Prevalência	0,2673	0,3317	0,4010
Taxa de Detecção	0,2327	0,1287	0,3564
Detecção de Prevalência	0,3119	0,1980	0,4901
Acurácia Balanceada	0,8811	0,6422	0,8329

classe para uma determinada amostra. Esta métrica indica 0,7460 para risco alto, 0,65 para risco médio e 0,7273 para risco baixo, mostrando que o modelo tem uma melhor capacidade de prever o diagnóstico de um paciente com risco alto. Por outro lado, o valor preditivo negativo indica a probabilidade de não pertencimento de uma amostra à uma classe quando o modelo prevê que a amostra não pertence à esta classe. Esta métrica indica 0,9496 para risco alto, 0,7469 para risco médio e 0,9126 para risco baixo. Este valor também mostra a capacidade do modelo de prever quando o paciente não é de alto risco. A taxa de detecção, que indica a incidência real da classe de risco, é de 0,2327 para risco alto, 0,3317 para risco médio e 0,4010 para risco baixo. A detecção de prevalência de 0,3119 em risco alto, 0,1980 em risco médio e 0,3564 em risco baixo. Por fim, a taxa de acurácia balanceada, que é a eficiência geral do modelo, calculada pela média aritmética da sensibilidade e especificidade, é de 0,8811 para risco alto, 0,6422 para risco médio e 0,8329 para risco baixo, mostrando a maior eficiência do modelo para classificar alto risco.

Triagem automática de gestantes de alto risco de mortalidade

4 Conclusões

Neste trabalho foi apresentada uma proposta de abordagem para triagem automática do fator de risco de saúde materna através da árvore de decisão. A base de dados era composta por dados públicos de diferentes hospitais, clínicas comunitárias das áreas rurais de Bangladesh, coletados por meio do sistema de monitoramento de risco baseado em IoT entre 2018 e 2020 [5]. A partir destes dados, através das análises estatísticas, foi possível concluir que os preditores BS e SystolicBP possuem maiores diferenças entre as diferentes classes de risco. Além disso, o modelo de classificação desenvolvido neste trabalho mostrou que BS e SystolicBP também podem ser usados como fatores de triagem para identificação de pacientes com alto risco de mortalidade materna. Os resultados mostraram que o modelo de árvores de decisão tem uma acurácia de 71,78% com uma sensibilidade de 87,04% para prever pacientes de alto risco de mortalidade.

Como trabalhos futuros, pretende-se melhorar o modelo utilizando modelos não lineares de redes neurais artificiais para investigação de sua capacidade preditiva através da acurácia, como também da sensibilidade.

References

1. Alam N., Mamun M., Dema P: Reproductive, Maternal, Newborn, Child, and Adolescent Health (RMNCAH): Key Global Public Health Agenda in SDG Era. In: Leal Filho W., Wall T., Azul A., Brandli L., Özuyar P. (eds) Good Health and Well-Being. Encyclopedia of the UN Sustainable Development Goals. Springer, Cham, Switzerland (2019) https://doi.org/10.1007/978-3-319-69627-0_52-1
2. Ahmed, M.; et al: Review and Analysis of Risk Factor of Maternal Health in Remote Area Using the Internet of Things (IoT). In: ECCE2019. pp. 357–365. Springer, Singapore (2020)
3. Sutton, R.T., et al.: An overview of clinical decision support systems: benefits, risks, and strategies for success. NPJ Digit. Med. **3**(17) (2020) <https://doi.org/10.1038/s41746-020-0221-y>
4. Indumathi, N., Ramalakshmi R., and Ajith, V.: Analysis of risk factors in the Firework Industries: Using Decision Tree Classifier. In: 2021 International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE). pp. 811–814. IEEE, Greater Noida, India (2021), <https://doi.org/10.1109/ICACITE51222.2021.9404726>
5. Ahmed, M.: Maternal Health Risk Data Set Data Set. Computer Science Department - Daffodil International University, Dhaka, Bangladesh, 2020. <https://archive.ics.uci.edu/ml/datasets/Maternal+Health+Risk+Data+Set> Acesso em : 3 Jun 2022.
6. Ochôa, F. Jimenez, et al.: Triagem automatizada de pacientes com risco de Câncer de Mama. In: Congreso Argentino de Informatica Y Salud (CAIS). pp. 49–58. Sadio, Buenos Aires, Argentina (2021)
7. Camargo, S. S., et al.: Modelagem Computacional de Indicadores Metabólicos para Estudo de Eficiência Reprodutiva em Vacas de Corte. In: X Congresso Brasileiro de Agroinformática (SBIAgro), pp. 857–866. Cbiagro, Ponta Grossa, Brasil (2015)

J. P. Martins; S. S. Camargo

8. Rochmawati, N., et al.: Covid Symptom Severity Using Decision Tree. In: 3rd International Conference on Vocational Education and Electrical Engineering (ICVEE), pp. 1-5. IEEE, Surabaya, Indonesia (2020). <https://doi.org/10.1109/ICVEE50212.2020.9243246>