AVALIAÇÃO DO DESEMPENHO DO ALGORITMO JRIP NA CLASSIFICAÇÃO DO DIAGNÓSTICO DE DOENÇAS CARDÍACAS

Ingrid Rafaella dos Santos Melo [1]
Natasha Seleidy Ramos de Medeiros[2]
Ingrid Bergmam do Nascimento Silva[3]
Larissa Duarte de Britto Lira[4]
Ronei Marcos de Moraes[5]

RESUMO

As técnicas de Mineração de Dados são utilizadas na identificação de informações relevantes em grandes volumes de dados para resolução de problemas reais. O WEKA é um dos softwares que contém diversas técnicas de mineração de dados que usam algoritmos que são utilizados para tomar de decisões. Este estudo tem como objetivo buscar o melhor resultado do algoritmo JRip no software WEKA na tentativa de distinguir a presença da doença cardíaca nos pacientes (valores 1,2,3,4) da ausência (valor 0) para o auxílio na tomada de decisão.

Palavras-chave: JRip, Algoritmo, WEKA, Desempenho.

INTRODUÇÃO

As doenças cardíacas constituem um grave problema de saúde pública no Mundo, sendo a principal causa de morte, de acordo com Organização Mundial de Saúde. Estima-se que 17,7 milhões de pessoas morreram por doenças do coração em 2015, representando 31% de todas as mortes em nível global (OMS, 2017).

Para realizar pesquisas mais avançadas alguns métodos computacionais vêm sendo utilizados para facilitar o processo de tomada de decisão sob um olhar científico, fornecendo

[1]Mestranda em Modelos de Decisão e Saúde da Universidade Federal da Paraíba- UFPB, ingridmeello@gmail.com;
[2]Mestranda em Modelos de Decisão e Saúde da Universidade Federal da Paraíba- UFPB, natashaseleidy@gmail.com;
[3]Mestranda em Modelos de Decisão e Saúde da Universidade Federal da Paraíba- UFPB, ingridgba2006@hotmail.com;
[4]Mestranda em Modelos de Decisão e Saúde da Universidade Federal da Paraíba- UFPB, larissadblira@hotmail.com;
[5]Professor da graduação e Pós-graduação no Centro de Ciências Exatas e da Natureza da Universidade Federal da Paraíba- UFPB, ronei@de.ufpb.br.

bons indicativos de precisão e padronização na análise dos dados(SOCZEK; ORLOVSKI, 2014).

A mineração de dados ou data mining surgiu da junção de três áreas cientificas que se relacionam são elas: Banco de dados, Estatística e Inteligência Artificial [KAWUU e CHUNG, 2015]. A mineração de dados é uma das etapas envolvidas no conceito do processo de descoberta de conhecimento (KDD) (Knowledge Discovery in Databases). O KDD é um processo que permite extrair conhecimento de informações armazenadas em grandes bases de dados especializadas. Conforme (figura 01). Neste estudo para aplicação de técnicas de mineração de dados foi utilizado o software WEKA (Waikato Environment for Knowledge Analysis). O software é um pacote de mineração de dados que possui vários algoritmos de classificação, esses algoritmos trabalham segundo a filosofia de aprendizagem de máquina, os quais operam para solucionar problemas de mineração de dados onde cada um deles possui padrões pré-definidos (NASERIPARSA; BIDGOLI; VARAE;2013).

No WEKA é possível executar tarefas que estão relacionada ao pré-processamento de dados como, por exemplo, a seleção e a transformação de atributos [WEKA, 2017].

Diante do exposto o objetivo do estudo é buscar o melhor resultado do algoritmo JRip para tentar diagnosticar a presença da doença cardíaca a partir das análises completas dos pacientes para assim auxiliar na tomada de decisão.

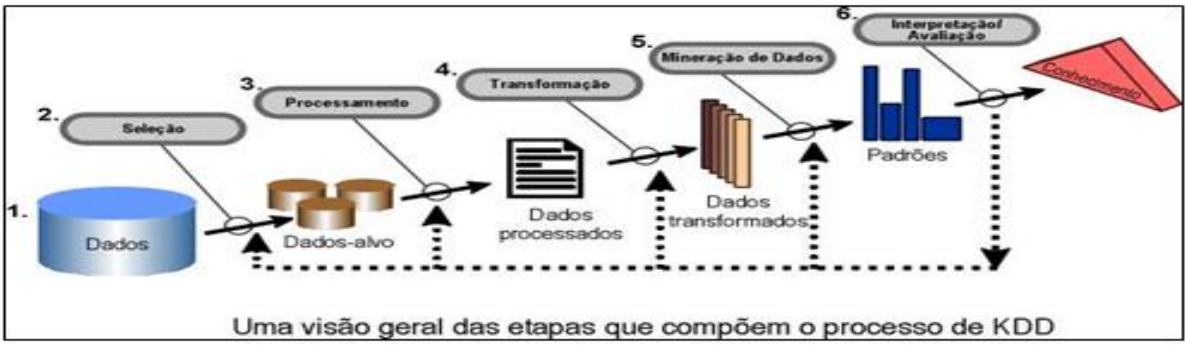

Figura 01: Etapas do processo de KDD (Fayyad et al., 1996) (adaptado).

METODOLOGIA

Modelo de decisão lógica clássica

Para tomar decisões é necessário utilizar métodos científicos a partir dos dados e/ ou informações. Estes podem ser baseados em: lógica (lógica clássica, lógica fuzzy, sistemas

especialistas), modelos (modelos probabilistas, modelos fuzzy, modelos em redes) e híbridos (dois ou mais sistemas utilizados para a mesma decisão) (PEREIRA et al., 2012).

Apesar da Lógica Clássica ser considerada praticamente o único sistema existente, até o século XX, sendo ainda hoje considerado um padrão de raciocínio correto. Seus princípios originados por Aristóteles descrevem que toda fórmula ou sua negação é verdadeira, não admitindo uma terceira condição (terceiro excluído). Sua fórmula e sua negação não podem ser ambas verdadeiras (não contradição) e que uma fórmula verdadeira é sempre verdadeira, e uma fórmula falsa é sempre falsa (identidade) (FASSBINDER, 2010).

Algoritmo JRip

Existem vários algoritmos que são utilizados para tomadas de decisão utilizando banco de dados extenso, porém há um modelo baseado em regras proposicionais ou condicionais, Repeated Incremental Pruning to Produce Error Reduction (RIPPER), ou seja este algoritmo realiza a poda incremental repetida para produzir redução de erro, também conhecido como JRip (indução baseada em regras), implementação em JAVA proposto por William W. Cohen (WITTEN; FRANK, 2005; FERREIRA, 2015; VERA, 2015).

Várias áreas utilizam o método RIPPER para tomadas de decisões, são elas: ciência, economia, enologia. biologia, saúde e outros. Alguns dos estudo interessantes foram: a detecção da quantidade de dióxido de enxofre na produção de vinhos (GOMEZ et al., 2017), definição de concessão de empréstimo bancário para pessoas jurídicas (Steiner, 2007), diagnóstico de pessoas com a doença de Alzheimer usando ao empregar o teste neuropsicológico (Shree; Sheshadri; Muralikrishna, 2016), distinguir casos de não-casos de asma em crianças (Afzal, 2013) e análise educacional (VERA, 2015; OLIVEIRA JÚNIOR, 2016).

O modelo é avaliado usando o classificador baseado em regras JRip, que implementa uma aprendizagem de regra proposicional, usa as regras representadas por IF - THEN (SE - ENTÃO). Uma regra IF - THEN é uma expressão do formulário, condição IF, então, conclusão. As regras são uma boa maneira de representar informação ou pedaços de conhecimento (Han et al., 2011; Cohen, 1995). O algoritmo basicamente divide-se em duas fases: a primeira gera um conjunto de regras para a comparação e a segunda otimiza o conjunto de regras iniciais para diminuir erros e tornar o processo mais seletivo, sendo esses passos repetidos inúmeras vezes no sistema WEKA (OLIVEIRA JÚNIOR, 2016).

O JRip implementa uma ordenação de classes seguindo a técnica “dividir-para-conquistar”, elencando linearmente o número de exemplos para treino (aprendizagem), realizando tal esquema para cada exemplo em sua base de regras. Isto é sequencialmente repetido até que as chances de erro sejam as menores possíveis de serem detectadas pelo sistema. A regra produzida com menor incidência de erro é eleita para a classificação. Ou seja, a classe que se sobressai é escolhida como padrão, auxiliando na determinação da classe minoritária (SOUZA, 2011).

DESENVOLVIMENTO

Neste estudo foi utilizado o banco de dados referente a doença cardíaca e os dados foram coletados na Fundação Clínica de Cleveland em 303 pacientes. O arquivo contém 76 atributos, porém todos os experimentos publicados referem-se ao uso de um subconjunto de 14 deles.

Este estudo utilizou um banco de dados disponível em repositórios internacionais (http://archive.ics.uci.edu/ml/datasets.html) sendo utilizado o banco Heart Disease em formato arff.

O banco de dados trata-se de um estudo do tipo experimental exploratório, transversal de abordagem quantitativa, onde a partir dele foram feitas realizações de experimentos e simulações de parametrização para a avaliação de desempenho do algoritmo Jrip no software WEKA na versão 3.8 para auxiliar na tomada decisão sobre o diagnóstico da doença do coração.As simulações realizadas foram feitas no *Cross-validation, Percentage Split* e no *Use training set.*

O atributo de decisão é o atributo Diagnóstico que se refere a presença ou a ausência da doença cardíaca no paciente, onde o campo de valor inteiro assume valores de intervalo de [0-4]. As análises completas dos pacientes levaram os médicos a distinguir, para cada um deles a presença de doença em níveis 1, 2, 3 e 4 da ausência valor 0.

A seguir a descrição completa dos 14 atributos:

	ATRIBUTOS
1	AGE
2	SEX
3	CP
4	TRESTBPS
5	CHOL
6	FBS
7	RESTECG
8	THALACH
9	EXANG

10	OLDPEAK
11	SLOPE
12	CA
13	THAL
14	NUM

Preparação de dados

Para o desenvolvimento deste artigo, foram utilizados registros históricos de um total de 303 paciente, dos quais 164 não apresentaram diagnósticos de doenças cardíacas, enquanto 139 apresentaram esses diagnósticos, conforme (tabela 03).

Foi realizada uma limpeza no banco que gerou uma redução de atributos, que passou de 76 para 14. Logo em seguida o arquivo foi transformado em arff para poder ser rodado no WEKA.

De cada um dos pacientes foi utilizado um subconjunto de 14 atributos, demonstrados detalhadamente abaixo:

- Idade em anos;
- Sexo: 0 = fêmea, 1 = macho;
- Tipo de doença: 1= angina típica, 2= angina atípica, 3 = sem angina, 4 = assintomático;
- Pressão do Sangue em repouso (mmHg na admissão ao hospital);
- Colesterol-S (mg/dl);
- Açúcar no sangue: > (mg/dl) (1=verdadeiro; 0=falso);
- Resultado do eletrocardiograma em repouso: 0 = normal; 1 = anomalia;
- Taxa máxima de batimentos cardíacos alcançada;
- Angina induzida pelo exercício: 0 = não, 1 = sim;
- ST depressão, induzida pelo exercício em relação ao repouso;
- Inclinação da rampa no exercício ST: 1 = inclinado para cima, 2 = em inclinação, 3 = inclinado para baixo;
- Número de artérias principais (0-3) coloridas pela fluoroscopia;
- Status do coração: 3= normal; 6 = problema permanente; 7 = problema reversível;
- Diagnóstico da doença de coração (angiographic disease status): 0=estreitamento < 50%, 1 = estreitamento > 50%.

Tabela 02 – Diagnóstico dos pacientes após completa avaliação médica (em níveis)

Banco de	Diagnóstico dos pacientes com estreitamento da artéria principal					
	0	1	2	3	4	
Cleveland	164	55	36	35	13	303

A tabela 03 mostra a nova configuração dos dados.

Tabela 03 – Configuração das Bases após a filtragem de dados

Banco de Dados	Diagnóstico dos Pacientes com Estreitamento da Artéria Principal		Total
	< 50%	> 50%	
Cleveland	164	139	303

A partir dos dados supracitados foram realizadas simulações no software WEKA, utilizando o JRip em busca do melhor desempenho do algoritmo para auxiliar na tomada de decisão do estudo.

RESULTADOS E DISCUSSÃO

Para obter os resultados, foram utilizados os parâmetros padronizados do algoritmo JRip retirados do próprio tutorial do WEKA.

Foram realizadas simulações no *Cross-validation* que tem como objetivo de construir um modelo baseado em subconjuntos dos dados fornecidos para calcular uma média e criar um modelo final. Foi utilizado também para realizar simulações a opção *Percentage split*, onde o WEKA toma um subconjunto percentual dos dados fornecidos para construir um modelo final. Outras simulações foram feitas também no *Use training set*, onde diz ao WEKA que para construir o modelo desejado, podemos simplesmente usar o conjunto de dados que fornecemos no arquivo ARFF. (ALBERNETHY, 2010). Após a realização das simulações, observamos que os resultados do *Cross-validation e* do *Percentage Split* não foram satisfatórios. O tipo de processamento que forneceu o melhor resultado foi o *Use training set* com KAPPA de 92,62%. Segue figura demonstrativa dos parâmetros utilizados no *Use*

training set (figura 2). Segue figura com os melhores resultados obtidos nas simulações do *Use training set*, *Cross-validation e* do *Percentage Split (*figura 4).

Figura 02: Parâmetros utilizados com melhor índice KAPPA.

```
=== Confusion Matrix ===

   a   b   c   d   e   <-- classified as
 164   0   0   0   0 |   a = 0
   5  50   0   0   0 |   b = 1
   4   0  32   0   0 |   c = 2
   3   0   0  32   0 |   d = 3
   2   0   0   0  11 |   e = 4
```

Figura 03: Matriz de confusão.

MELHORES RESULTADOS		
USER TRAINING	KAPPA	0.9262
PERCENTAGE SPLIT	KAPPA	0.2387
CROSS VALIDATION	KAPPA	0.2098

Figura 04: Melhores resultados.

Na Matriz de confusão do melhor resultado (Figura 03) podemos observar que o modelo é adequado. Foi detectado na sua diagonal principal 95,38% de precisão de acertos. Em questão de erros, verifica-se a ocorrência de 14 Falsos Positivos no qual o indivíduo sadio foi dado como doente, 4,62%. Para obter esse resultado foi necessário retirar a poda incremental em razão de obter uma regra perfeita e condição satisfatória, também se fez necessário simulações de retirada de variável. Mediante isso foi constatado que com a retirada da variável de número 4 [trestbps: pressão arterial em repouso (em mmHg na admissão ao hospital)] o KAPPA alcançou seu melhor resultado.

O coeficiente de concordância de Kappa, sugerido por Cohen em 1960 tem a finalidade de medir o grau de concordância entre proporções derivadas de amostras dependentes (SOUZA; PAES, 2012).

	REGRAS	RESULTADO	ACERTOS	ERROS
1	SE (oldpeak >= 2.4) and (oldpeak <= 2.6) and (isto = 7) and (age >= 50) and (age <= 58)	ENTÃO num=4	2	0
2	SE (ca = 3) and (chol >= 289) and (thalach >= 124) and (exang = 0)	ENTÃO num=4	3	0
3	SE (oldpeak >= 2) and (chol <= 212) and (age >= 60) and (thalach <= 132)	ENTÃO num=4	4	0
4	SE (ca = 3) and (age >= 65) and (oldpeak <= 1) and (cp = 4)	ENTÃO num=4	2	0
5	SE (isto = 7) and (thalach <= 139) and (oldpeak >= 2) and (age <= 56) and (chol <= 239)	ENTÃO num=3	6	0
6	SE (ca = 2) and (age <= 59) and (oldpeak >= 3)	ENTÃO num=3	4	0
7	SE (isto = 7) and (thalach <= 132) and (chol >= 269) and (oldpeak >= 2)	ENTÃO num=3	3	0
8	SE (cp = 4) and (ca = 2) and (oldpeak <= 1) and (fbs = 1)	ENTÃO num=3	4	0
9	SE (cp = 4) and (thalach <= 132) and (oldpeak <= 1.2) and (oldpeak >= 0.8) and (thalach >= 125)	ENTÃO num=3	3	0

10	SE (isto = 7) and (ca = 3) and (chol >= 309)	ENTÃO num=3	2	0
11	SE (thalach <= 115) and (ca = 1) and (isto = 3)	ENTÃO num=3	2	0
12	SE (isto = 7) and (chol <= 164) and (cp = 4)	ENTÃO num=3	2	0
13	SE (ca = 2) and (thalach >= 152) and (thalach <= 157)	ENTÃO num=3	2	0
14	SE (isto = 7) and (chol >= 270) and (chol <= 281) and (restecg = 2) and (oldpeak <= 1.6)	ENTÃO num=3	4	0
15	SE (thalach <= 147) and (cp = 4) and (oldpeak >= 3)	ENTÃO num=2	7	0
16	SE (cp = 4) and (oldpeak >= 1) and (chol >= 282) and (chol <= 294)	ENTÃO num=2	5	0
17	SE (thalach <= 147) and (isto = 6) and (fbs = 1)	ENTÃO num=2	3	0
18	SE (cp = 4) and (ca = 1) and (age >= 58) and (thalach >= 147)	ENTÃO num=2	4	0
19	SE (thalach <= 146) and (oldpeak >= 1.2) and (chol >= 234) and (chol <= 263) and (age >= 61)	ENTÃO num=2	4	0
20	SE (exang = 1) and (chol >= 305) and (age <= 56)	ENTÃO num=2	4	0
21	SE (cp = 4) and (chol <= 239) and (oldpeak >= 1) and (oldpeak <= 1.4) and (age >= 59)	ENTÃO num=2	3	0
22	SE (thalach <= 156) and (chol <= 212) and (ca = 1) and (oldpeak <= 1.4)	ENTÃO num=2	2	0
23	SE (isto = 7) and (cp = 4) and (oldpeak >= 1.4)	ENTÃO num=1	11	0
24	SE (cp = 4) and (exang = 1) and (chol >= 244) and (oldpeak >= 1)	ENTÃO num=1	6	0
25	SE (isto = 7) and (chol >= 241) and (chol <= 255)	ENTÃO num=1	4	0
26	SE (age >= 57) and (sex = 1) and (chol >= 273) and (age <= 65)	ENTÃO num=1	8	0
27	SE (isto = 7) and (age <= 50) and (chol >= 223) and (thalach <= 168)	ENTÃO num=1	4	0
28	SE (age >= 56) and (age <= 61) and (restecg = 2) and (ca = 1) and (chol <= 236)	ENTÃO num=1	4	0
29	SE (cp = 4) and (oldpeak <= 0) and (chol <= 197)	ENTÃO num=1	3	0
30	SE (age >= 59) and (age <= 61) and (cp = 4) and (sex = 0)	ENTÃO num=1	3	0
31	SE (sex = 1) and (thalach <= 162) and (oldpeak >= 2.4) and (age >= 60)	ENTÃO num=1	2	0
32	SE (sex = 1) and (thalach <= 162) and (ca = 2)	ENTÃO num=1	2	0
33	SE (sex = 1) and (thalach <= 152) and (cp = 3) and (age <= 49) and (age >= 46)	ENTÃO num=1	3	0
34		ENTÃO num=0	178	14

CONSIDERAÇÕES FINAIS

Observou-se que o modelo JRip mostrou de grande valia ao ser utilizado para identificar o diagnóstico da doença cardíaca através de análises em pacientes. As regras serviram para nos mostrar quais eram os fatores dentre os atributos presentes nesse estudo que influenciaram na decisão de ter ou não ter a doença cardíaca. As pessoas podem efetivamente alcançar o diagnóstico correto por causa das vantagens do algoritmo, tanto os pacientes quanto os médicos podem obter enormes benefícios usando este sistema de diagnóstico. O JRip pode ser utilizado por outras áreas para produção de redução de erros em determinados estudos através de regras e poda incremental repetida.

REFERÊNCIAS

ABERNETHY, M et al. Mineração de dados com WEKA, Parte 1: Introdução e regressão: **Developer Works**. 2010

EDEKI, C. PANDYA, S. Comparison of Data Mining Techniques used to Predict Cancer Survivability.

International Journal of Computer Science and Information Security, 10(6): 1-6, 2012.

FASSBINDER, A. G. O. Uma Apresentação dos Principais Sistemas Relacionados à Lógica Clássica. 2010. 307p. Dissertação (mestrado). **Universidade Federal de Santa Catarina. Programa de Pós-Graduação em Ciência da Computação**. Florianópolis - SC, 2010.

FERREIRA, J. A. et al. Assessment of fuzzy gaussian naive bayes for classification tasks.In: **The Seventh International Conferences on Pervasive Patterns and** Applications. [S.l.: s.n.], 2015.

KAWUU e CHUNG, 2015. kawuu W. Lin, Sheng-Hao Chung. A fast and resource efficient mining algorithm for discovering frequent patterns in distributed computing environments. **Future Generation Computer Systems**. Elsevier Journal 52. 2015.

NASERIPASA, M.; BIDGOLI, A.; VARAEE, T. Improving Performance of a Group of Classification Algorithms Using Resampling and Feature Selection. **World of Computer Science and Information Technology Journal**. 3(4): 70-76, 2013.

OLIVEIRA JÚNIOR, J.G. **Pattern Identification for Dropout Analysis in Undergraduate Courses using Educational Data Mining.** 2015. 86 f. Dissertação - Programa de Pós-graduação em Computação Aplicada, Universidade Tecnológica Federal do Paraná. Curitiba, 2015.

ORGANIZAÇÃO MUNDIAL DA SAÚDE (OMS). **Doenças Cardiovasculares, (2017).** Disponível em: <https://www.paho.org/bra/index.php?option=com_content&view=article&id=5253:doencas-cardiovasculares&Itemid=839 >. Acesso em: 18/09/2018.

ORLOVSKI, R. Mineração de Dados: Conceitos e aplicação de algoritmos em uma Base de Dados na área da saúde. **Revista Científica Semana Acadêmica**, v. 01, p. 01, 2014.

PEREIRA, et al. Condições de Acesso às Pessoas com Deficiência em Instituições de Ensino Enfermagem: Utilização de Redes Neurais Artificiais como Suporte à Decisão. **Revista Brasileira de Ciências da Saúde.** v. 16, n. 2, p. 143-148, 2012.

Roth GA, Forouzanfar MH, Moran AE, Barber R, Nguyen G, Feigin VL, et al. Demographic and epidemiologic drivers of global cardiovascular mortality. **N Engl J Med**. 2015 Apr; 372:1333-41.

SOCZEK, F. C.; ORLOVSKI, R. Mineração de Dados: Conceitos e aplicação de algoritmos em uma Base de Dados na área da saúde. **Revista Científica Semana Acadêmica**, v. 01, p. 01, 2014.

SOUZA; PAES **Teste de concordância Kappa:** Por dentro da estatística. Setor de Estatística Aplicada, Pró-Reitoria de Pós-Graduação e Pesquisa, Universidade Federal de São Paulo – UNIFESP, São Paulo (SP), Brasil 2012.

VARAEE, T. Improving Performance of a Group of Classification Algorithms Using Resampling and Feature Selection. **World of Computer Science and Information Technology Journal**. 3(4): 70-76, 2013.

VERA, C.M. Prediccion del Fracaso y el Abandono Escolar Mediante Técnicas de Minería de Datos. Universidad de Córdoba. Tesis doctoral, 2015. Disponível em: http://www.uco.es/grupos/kdis/docs/thesis/2015-CMarquez.pdf. Acesso em 13/09/2018.

WEKA, 2017, UNIVERSITY OF WAIKATO. Weka 3.8 – Machine Learning Software in Java. Disponível no site da **University of Waikato** (2017). URL: http://www.cs.waikato.ac.nz/ml/weka/ downloading.html.

WITTEN, I.H.; FRANK, E. Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations. **Morgan Kaufmann Publishers**, 2. ed. San Francisco, California, 2005.