

## APLICAÇÃO DO MÉTODO JRIP EM VARIADOS BANCOS DE DADOS

Ana Carolina Costa de Oliveira<sup>1</sup>  
Emanuelle Silva de Mélo<sup>2</sup>  
Thayana Rose de Araújo Dantas<sup>3</sup>  
Ronei Marcos de Moraes<sup>4</sup>  
Messias Rafael Batista<sup>5</sup>  
Marcelo Fernandes de Sousa<sup>6</sup>

### RESUMO

**Introdução:** o algoritmo JRip é um método de classificação que utiliza poda incremental reduzida de erro, para proceder ao tratamento de todas as instâncias e se chegar a uma tomada de decisão baseando-se em regras, que, por sua vez, induz os dados de treinamento através de sequências lógicas. **Objetivo:** Analisar o comportamento do algoritmo JRip em bancos de dados com diferentes características. **Metodologia:** uma investigação do tipo descritiva exploratória, na qual foram utilizados os Algoritmos de Fragmentação do *Weka-Cross-Validation (Folds)* e *Percentage Split* em dez bases de dados, a partir de critérios de classificação pré-estabelecidos para o processamento dos dados. **Resultados:** Por meio dessas rotas de análise, percebeu-se que o JRip obteve concordância quase perfeita com 50% dos bancos de dados. Desses, os que tiveram melhor taxa de acerto foram o *Hypothyroid* (99.34%), *Nursery* (96.84%) e *Vote* (95.63%). Os melhores resultados foram obtidos após o processamento dos dados com o *Cross-Validation*, enquanto, no *Percentage Split*, os resultados foram melhorando à medida que se aumentava o percentual. Mas, ainda assim, as alterações não foram satisfatórias para superar os resultados obtidos pelo *Cross-Validoatin*. **Conclusões:** preconiza que o algoritmo JRip é um bom classificador para bancos da área da saúde. Porém, mais estudos envolvendo o método isolado precisam ser realizados para garantir a sua eficácia em múltiplos bancos de dados.

**Palavras-chave:** Banco de dados; Classificador; JRip; WEKA.

### ABSTRACT

**Introduction:** The JRip algorithm is a classification method that uses reduced incremental error pruning, to proceed with the treatment of all instances and arrive at a decision-making based on rules, which, in turn, induces the data of training through logical sequences. **Objective:** To analyze the behavior of the JRip algorithm in databases with different characteristics. **Methodology:** an exploratory descriptive investigation, in which the *Weka-Cross-Validation (Folds)* and *Percentage Split Fragmentation Algorithms* were used in ten databases, based on pre-established classification criteria for data processing. **Results:** Through these analysis routes, it was noticed that JRip obtained almost perfect agreement with 50% of the databases. Of these, those with the best hit rate were *Hypothyroid* (99.34%), *Nursery* (96.84%) and *Vote* (95.63%). The best results were obtained after processing the data with *Cross-Validation*, while in the *Percentage Split*, the results improved as the percentage

<sup>1</sup> UNIESP. E-mail: prof2121@iesp.edu.br

<sup>2</sup> UFPB. E-mail: emanuelleacupunturista@gmail.com

<sup>3</sup> UFPB. E-mail: taiane\_camel@yahoo.com.br

<sup>4</sup> UFPB. E-mail: ronei@de.ufpb.br

<sup>5</sup> UNIESP. E-mail: messias.rafael@iesp.edu.br

<sup>6</sup> UNIESP:E-mail: marcelo@iesp.edu.br

increased. But, even so, the changes were not satisfactory to overcome the results obtained by Cross-Validoatin. Conclusions: recommends that the JRip algorithm is a good classifier for healthcare banks. However, more studies involving the isolated method need to be performed to ensure its effectiveness in multiple databases.

**Keywords:** Database; Classifier; JRip; WEKA

## 1 INTRODUÇÃO

A busca constante de informação em diferentes áreas do conhecimento suscita a necessidade de ferramentas que auxiliem a tomada de decisão e a busca por informações corretas e seguras (SOCZEK; ORLOVSKI, 2014; DOUBRAVSKY; DOHNAL, 2015). Com o propósito de aperfeiçoar o processo de tomada de decisão, novas formas de geração de conhecimento vêm sendo utilizados, por exemplo, os modelos matemáticos e computacionais.

A utilização de algoritmos computacionais robustos e eficientes são uma alternativa para desenvolver um modelo de tomada de decisão eficaz (CHRISTOPHER *et al.*, 2015). Esses algoritmos funcionam baseados em conjunto de regras de classificação, sistema especialistas *fuzzy*, redes neurais artificiais, agentes inteligentes entre outros (LINS; WEBBER, 2019).

Para os algoritmos baseados em conjunto de regras de classificação, tem-se a lógica clássica, também conhecida por lógica formal, como ciência que estuda os critérios de validação, condutores do pensamento e da demonstração, isto é, ciência do raciocínio dedutivo (FORBELLONE; EBERSPÄCHER, 2000).

A lógica clássica trata basicamente dos chamados conectivos lógicos de negação, conjunção, disjunção, implicação e bicondicional, sobre os quantificadores existencial e universal e sobre o predicado de igualdade e suas variações (D’OTTIAVIANO; FEITOSA, 2009). Uma limitação desse tipo de sistema pode estar relacionada ao tamanho da amostra e quantidade de dados complexos, podendo contribuir para a pobreza dos resultados (COHEN, 1995).

Desse modo, com o intuito de promover uma melhor leitura dos dados gerados pela lógica clássica, segundo Govada, (2016), foi criado o algoritmo *Repeated Incremental Pruning to Produce Error Reduction* (RIPPER) ou Poda Incremental Repetida para Produzir Redução de Erro, proposto por William Cohen em 1995 pela Universidade de Waikato na Nova Zelândia, como uma versão otimizada do *Induction Reduced Error Pruning* (IREP). O RIPPER utiliza um conjunto de regras de classificação que dão origem a um modelo simplificado de forma interativa através da poda incremental, repetida para redução do erro (COHEN, 1995; BOUCKAERT *et al.*, 2010; KSHIRSAGAR; JOSHI, 2016).

O algoritmo apresenta uma versão na linguagem Java, o JRip, presente na ferramenta tecnológica *Waikato Environment for Knowledge Analysis* (WEKA), composta de um conjunto de algoritmos de aprendizado de máquina para tarefas de mineração de dados (HALL *et al.*; 2009). Essa tecnologia apresenta um código aberto, além de possuir vários algoritmos de modelos de decisão (J48, Fúria (*Fuzzy*), *DecisionTable*, *Naive Bayes*, entre outros) e regressão (BOUCKAERT *et al.*; 2010).

Dentre os métodos de classificação do WEKA, o JRip é um dos mais básicos e populares, tendo como finalidade a construção de um conjunto de regras que moldam um conjunto de dados (RAJPUT *et al.*, 2011; WASEEM *et al.*, 2013; FAIZAL *et al.*, 2018).

O JRip utiliza poda incremental reduzida de erro para proceder ao tratamento de todos os registros, a fim de se chegar a uma tomada de decisão ou formação de classes, baseando-se em regras, que induzam os dados de treinamento por meio de uma sequência do tipo SE <condição> E < condição > E ... ENTÃO<classe> (BORGES, 2009; RAJPUT *et al.*, 2011;

PARSANIA *et al.*, 2014). De acordo com Oliveira Júnior (20016), algoritmo é dividido em 2 fases, sendo a 1<sup>o</sup> fase responsável por gerar um conjunto de regras para a comparação e a 2<sup>a</sup> fase por otimizar um conjunto de regras iniciais para diminuir erros, tornando assim, o processo mais seletivo. Segundo Melo *et al.* (2018) o JRip utiliza a ordenação de classes a partir da técnica “dividir-para-conquistar”, em que o número de treinamento (aprendizagem) é repetido até que as chances de erro sejam as menores possíveis de serem detectadas pelo sistema. Enquanto, as regras com menor incidência de erro é definida para a classificação.

Já para um problema que envolva várias classes, elas são ordenadas de maneira crescente em relação ao número de padrões obtidos dentro de cada uma. A primeira classe, a menor, é considerada positiva e as demais negativas. O mesmo processo é repetido com todas as classes, seguindo a ordem de classificação, selecionando o melhor subconjunto ao término do processo (BORGES, 2009; SOUZA, 2011; PARSANIA *et al.*, 2014).

A maior efetividade do algoritmo está relacionada à formação de um conjunto de regras que cubram todos os membros de cada classe. Um conjunto de regras inicial é gerado utilizando uma variante proveniente de uma regra vazia e outra da adição de antecedentes da regra original; se ainda houver pontos positivos residuais, mais regras serão geradas (WITTEN *et al.*, 2016; BORGES, 2009; PARSANIA *et al.*, 2014). Ante o exposto, na busca por explorar a funcionalidade e efetividade do JRip, o presente estudo tem como objetivo analisar o comportamento do algoritmo JRip em bancos de dados com diferentes características.

## 2 METODOLOGIA

Este item deve conter o tipo de pesquisa, a forma de abordagem, procedimentos e modalidade da pesquisa.

A pesquisa apresentada trata-se de um estudo do tipo descritivo e de caráter exploratório. O método da Lógica Clássica foi utilizado através da implementação do algoritmo JRip, presente no *software* Weka, com distribuição *Open Source*, versão 3.6. Uma importante característica a ser ressaltada é como o banco de teste é definido.

O WEKA dispõe de quatro opções de teste, a saber: dados de treinamento (quando os dados de treinamento são utilizados tanto na criação do modelo quanto para o teste), dados de teste (quando um arquivo separado de teste é carregado com esta finalidade), *cross-validation* e *percentage split*. Visto que os bancos de dados utilizados neste trabalho não possuem bancos de testes disponíveis, apenas as três últimas opções ficam disponíveis.

### 2.1 CARACTERIZAÇÃO DOS BANCOS

Para analisar o comportamento do JRip, foram utilizados dez bancos de dados: *arrhythmia*, *breast-cancer*, *dermatology*, *diabetes*, *ecoli*, *hypothyroid*, *nursery*, *postoperativepatientdata*, *vote* e *zoo*, disponíveis em repositórios internacionais, por meio do link: <https://archive.ics.uci.edu/ml/datasets.php>. A seguir, serão apresentadas as características de cada banco de dados:

BANCO DE DADOS	FINALIDADE	NÚMERO DE INSTÂNCIAS	NÚMERO DE ATRIBUTOS	NÚMEROS DE CLASSES	FONTE
<i>Arrhythmia</i>	Determinar o tipo de arritmia, a partir dos resultados do exame eletrocardiograma.	452	279	16	Guvenir <i>et al.</i> , 1997.

<b>Breast-cancer</b>	Determinar a ocorrência do câncer de mama na Iugoslávia, região dos Balcãs.	285	9	2	Zwitter e Soklic, 1988.
<b>Dermatology</b>	Realizar de diagnóstico diferencial das doenças eritemato-escamosas.	366	34	6	Güvenir <i>et al.</i> , 1998.
<b>Diabetes</b>	Prever o aparecimento de diabetes mellitus em um grupo indígena.	768	8	2	Smith <i>et al.</i> , 1988.
<b>Ecoli</b>	criar um sistema probabilístico de posicionamento das proteínas em células eucarióticas.	336	8	8	Nakai e Kanehisa, 1992.
<b>Hypothyroid</b>	Divulgar registros da doença de tiróide, fornecidos pelo Instituto de Pesquisa Médica, de Sidney, Austrália.	3772	30	5	Garavan e Quinlan, 1987.
<b>Nursery</b>	Classificar pedidos de creches.	12960	8	5	Bohanec e Rajkovic, 1990.
<b>Postoperative-patient-data</b>	Determinar para onde os pacientes em uma área de recuperação pós-operatória devem ser enviados, uma vez que, a hipotermia é uma preocupação significativa após a cirurgia.	90	9	3	Woolery <i>et al.</i> , 1991.
<b>Vote</b>	Incluir votos de cada representante congressista da Casa dos Estados Unidos.	435	17	2	Schlimmer, 1987.
<b>Zoo</b>	Classificar variados grupos de animais conforme suas características específicas.	101	18	7	Richard, 1990.

## 2.2 PROCESSAMENTO E ANÁLISE DOS DADOS

Para o processamento dos dados, utilizou-se os algoritmos de fragmentação ofertados pelo Weka: *Cross-Validation (folds)* e *Percentage Split*, que tratam da interação do programa com cada banco de dados.

Inicialmente, aplicou-se o *Cross-Validation*, que realiza repetição de interações, cujos *folds* são os números de pares e subconjuntos de treinamento, fornecidos na entrada.

A partir de então, são gerados os dados estatísticos, apresentando as informações. Como os bancos de dados empregados possuem números variados de instâncias para este estudo, a quantidade de instâncias foi determinante para uma padronização quanto à escolha dos *folds*, conforme apresentado na tabela 1.

Tabela 1. Padronização dos *folds* segundo a quantidade de instâncias

N. de Instâncias	Cross-Validadtion Folds
<100	20
100-500	15
>500	10

Fonte: Autores

Em seguida, foi utilizado o processo de fragmentação, o *Percentage Split*, para verificar o comportamento dos resultados ao utilizar o algoritmo JRip.

Nessa etapa, um subgrupo de treinamento é formado com x% do tamanho do banco de dados, sendo o valor de x a percentagem dada, por meio da qual obtém-se a redução do tamanho dos bancos de dados com o intuito de proporcionar melhores resultados estatísticos. Os percentuais utilizados para cada banco de dados foram: 10, 20 e 30%.

O JRip apresenta alguns parâmetros de configuração para a sua execução. Desse modo, como preconizados, utilizou-se os seguintes critérios: *folds*- quantidade de dados utilizados para a poda (F=3); *minNo* - peso mínimo total das instâncias em uma regra (N=2); *optimizations* – número para otimização, que ocorre após a produção das regras e a fase de poda (O=2); *seed*- semente para randomização dos dados (S=1) (Cohen,1995; Rajput et al., 2011; Parsania et al., 2014).

No que se refere ao classificador, as regras precisam ser avaliadas com relação a sua pertinência e sua funcionalidade, averiguando a capacidade de representação dos conhecimentos advindos dos dados. Nessa perspectiva, alguns parâmetros foram utilizados para averiguar a concordância do método com os bancos: a taxa de acerto do método, o índice Kappa (*k*) e a matriz de confusão.

A taxa de acerto do método implica na percentagem geral de quanto o algoritmo conseguiu classificações corretas, a partir do número total de instâncias (Romão, 2002). Já a matriz de confusão é usada para mensurar a precisão da classificação do modelo, uma forma simples de se medir o percentual de decisões corretas, partindo da soma dos valores na diagonal principal da matriz, dividida pela soma de todos os valores da matriz (Manimaran & Vanitha, 2017). E por fim, há o índice Kappa (Sheskin, 2003).

Tabela 2. Interpretação dos valores de Kappa (K)

Valor de Kappa(k)	Interpretação
<0	Nenhuma concordância
0-0.20	Leve concordância
0.21-0.40	Concordância regular
0.41-0.60	Concordância moderada
0.61-0.80	Concordância substancial
0.81-1	Concordância quase perfeita

Fonte: Sheskin (2003)

Na tabela 2, encontra-se a interpretação dos valores de Kappa adotados no estudo.

### 3 RESULTADO E DISCUSSÃO

De acordo com os critérios de classificação para o processamento dos dados pelo *cross-validation*, os bancos escolhidos para cada percentual de *folds* preestabelecido estão apresentados na tabela 3.

Tabela 3. Classificação dos bancos para processamento utilizando o *cross-validation*, segundo o número de instâncias

Cross-Validation ( <i>Folds</i> )	Bancos de Dados
20 (<100)	Postoperative-paciente-data
15 (100-500)	Arrhythmia, Breast-cancer, Dermatology, Ecoly, Vote e Zoo
10 (>500)	Diabetes, Hypothyroid, Nursery

Fonte: Autores

Observa-se que os bancos com maior taxa de acerto em relação à classificação das instâncias (tabela 4) foram *Hypothyroid*, *Nursery* e *Vote*, com 99,34%, 96,84% e 95,63%, respectivamente. Estes valores corroboram com o nível de concordância entre os bancos e o método JRip, de modo que, os índices Kappa também foram maiores nesses três bancos (0.955, 0.954 e 0.909, respectivamente).

Tabela 4. Caracterização dos bancos após processamento pelo JRip

Banco	Regras	Taxa de Acerto (%)	Índice Kappa	Tempo de CPU
Arrhythmia	13	69.25	0.5234	1.45 <i>seconds</i>
Breast-c&er	3	66.78	0.1562	0.04 <i>seconds</i>
Dermatology	15	87.70	0.8461	0.04 <i>seconds</i>
Diabetes	4	76.04	0.4538	0.31 <i>seconds</i>
Ecoly	10	82.14	0.7496	0.31 <i>seconds</i>
Hypothyroid	5	99.34	0.9549	2.28 <i>seconds</i>
Nursery	131	96.84	0.9536	115.92 <i>seconds</i>
Postoperative-paciente-data	2	71.11	0	0.01 <i>seconds</i>
Vote	4	95.63	0.9088	0.03 <i>seconds</i>
Zoo	6	87.13	0.8239	0 <i>seconds</i>

Fonte: Autores

Em contrapartida, o banco *Postoperative-patientdata* obteve o pior índice Kappa (0), apesar de esta medida não estar associada simultaneamente a uma pior taxa de acerto.

A percentagem de acerto implica na taxa geral de quanto o algoritmo conseguiu classificações corretas a partir do número total de instâncias. Logo, acreditasse que essa medida é a mais confiável, por ser representativa do total de dados investigados (Romão, 2002).

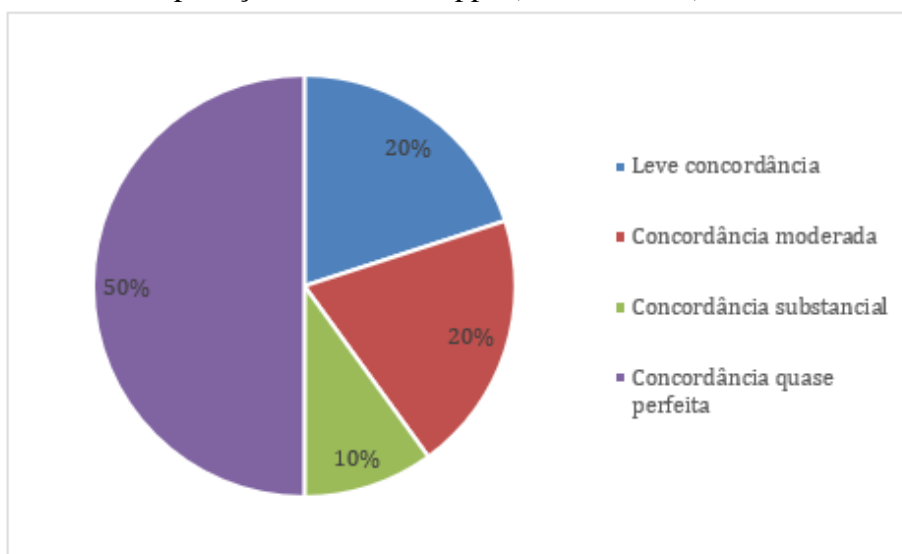
Observou-se que o JRip obteve bons valores para taxa de acerto e de concordância, em estudos realizados para medir o desempenho de estudantes de Programação Java, sendo o Índice Kappa também uma medida de confiança (TARUN *et al.*, 2014; EVALE *et al.*, 2016).

Conforme demonstra a tabela 4, o banco com maior número de regras produzidas foi o *Nursery*, com 131 regras. Porém, a quantidade dessas regras não foi determinante na concordância entre os bancos e o método. Em contrapartida, o banco *Hypothyroid*, com apenas 5 regras geradas, obteve o melhor índice Kappa (0.955).

A estatística Kappa é um índice que mensura o nível da concordância e ligação entre os dados. Seus resultados podem variar de 0 a 1, levando em consideração que, quanto mais próximo de 1, maior é a confiança dos resultados (SHESKIN; 2003).

De acordo com a classificação do índice Kappa, 80% dos bancos obtiveram concordância moderada quase perfeita (Kappa 0,45-0,95) pelos critérios estabelecidos, Sendo 50% dos bancos de dados com concordância quase perfeita, 10% concordância substancial e 20% concordância moderada, conforme figura 1.

Figura 1: Grau de concordância entre o JRip e os bancos, seguindo critérios de interpretação do Índice Kappa (Sheskin 2003).



Fonte: Autores

Em relação ao processamento dos dados pela CPU (Tabela 4), o banco *Nursery*, com 12960 instâncias, levou maior tempo para realizar as análises (115.92 segundos), seguido do *Hypothyroid* (3772 instâncias), que levou 2,28 segundos.

O JRip tem demonstrado levar um tempo maior na geração dos dados, quando comparado a outros algoritmos, como por exemplo J48 (MANIMARAN; VANITHA, 2017).

Por sua vez, a Matriz de Confusão indica a quantidade de erros e acertos das classificações das instâncias, sendo utilizada para medir a sensibilidade, a especificidade e a precisão do teste, representando, portanto, a autenticidade dos resultados (MANIMARAN; VANITHA, 2017). Nesse sentido, todavia, pôde-se observar erros em bancos da área de saúde, o que pode implicar em comprometimentos irreversíveis envolvendo seres humanos (Tabela 5 e Tabela 6).

Tabela 5. Matriz de confusão do banco *Postoperative-paciente-data*

a	b	c	Classe
604	0	0	a=ambulatório

2	0	0	b= UTI
24	0	0	c=casa

Fonte: Autores

Tabela 6. Matriz de confusão do banco *Diabetes*

a	b	Classe
286	69	a=negativo
70	113	b= positivo

Fonte: Autores

Observa-se que, em relação ao banco *postoperativepatient-data*, os dois únicos pacientes, que deveriam ser encaminhados para a UTI, foram classificados no grupo de encaminhamento para ambulatório hospitalar. O mesmo ocorreu com todos os pacientes que deveriam receber alta.

Implica dizer que, no primeiro caso, os pacientes correriam risco de vida por não terem assistência intensiva adequada, enquanto, no segundo caso, os pacientes, que deveriam estar em casa, estão ocupando os leitos ambulatoriais, que poderiam ser ocupados por outros pacientes realmente necessitados.

Em relação ao estudo envolvendo o diagnóstico de diabetes, obtém-se com o teste que dos 183 pacientes que possuem a doença, apenas 113 tiveram confirmação por exame. Ou seja, 70 indivíduos com diabetes não saberiam que possuem a patologia, o que implicaria na ausência de cuidados fundamentais para o controle da doença (KUMAR; VELIDE, 2014). Em ambas as situações, pode-se concluir que o método JRip não foi um bom classificador.

No estudo de Manimaran e Vanitha (2017), apesar de ter alcançado o valor de 85% de precisão nos resultados envolvendo dados para o diagnóstico da diabetes, o JRip obteve baixo desempenho na determinação dos acertos. No entanto, no estudo realizado por Lakmali e Haddela (2017), o algoritmo foi um bom classificador, porém não envolvia dados da área de saúde.

Ao processar os dados utilizando o *Percentage Split*, pôde-se observar que os resultados referentes ao índice Kappa foram melhorando à medida que se aumentava o percentual, embora ainda não tenha suficiente para superar os resultados obtidos pelo *Cross-Validation*.

Não corroborando com o estudo Soczek e Orlovski (2014), no qual, ao utilizar o JRip pelo processamento Split como um dos algoritmos de regras de classificação, obteve-se resultados na taxa de acerto e estatística Kappa em bancos voltados para a área da saúde.

#### 4 CONSIDERAÇÕES FINAIS

O algoritmo JRip foi um bom classificador em 50% dos bancos, com que apresentou concordância quase perfeita. Nesse sentido, como os bancos de dados, em sua maioria, são representantes da área de saúde, pôde-se verificar que o JRip foi um bom método de classificação para se trabalhar com dados dessa natureza, apesar de não ter obtido bons resultados frente à especificidade, sensibilidade e precisão do teste, quando aplicado aos bancos *postoperativepatient-data* e *diabetes*, de acordo com as matrizes de confusão.

Entretanto, é preciso se atentar às diversas medidas envolvidas entre métodos de classificação e os diversos bancos de dados, especialmente, quando estes envolvem decisões



na área da saúde, uma vez que, resultados insatisfatórios podem gerar danos irreversíveis, envolvendo seres humanos.

Com relação às medidas de classificação do JRip, não se observou relação entre as taxas de acerto com o número de regras, o tempo de processamento da CPU, bem como com o índice Kappa. Cabe, portanto, ao pesquisador, mais uma vez, analisar a pertinência da aplicabilidade do método para a tomada de decisões envolvendo os bancos de dados de interesse.

Pode-se concluir que os melhores resultados para taxas de acerto e índice Kappa foram obtidos por meio do *cross-validation*, quando comparados aos resultados provenientes do processamento com o *percentage split*, mesmo utilizando o percentual máximo pré-definido (30%).

Assim, o JRip torna-se um algoritmo com restrições por possuir regras de classificação padronizadas, fato que dificulta a sua exploração para a obtenção de resultados em variados bancos de dados.

Pôde-se observar ainda uma escassez de estudos, envolvendo a utilização do método na literatura, sobretudo, que apresentassem a aplicação do JRip, de maneira isolada, a bancos de dados. Esse foi um fator que tornou a nossa discussão, quanto a aplicabilidade às diferentes naturezas dos dados, de certa forma, limitada.

Sugere-se, portanto, que mais estudos sejam fomentados, a partir do algoritmo JRip em variados bancos de dados, de modo que possam contribuir com conclusões mais fidedignas acerca do seu comportamento, em estudos, que envolvam diferentes áreas de conhecimento.

Diante disso, este estudo torna-se relevante no sentido de contribuir para o avanço de novas pesquisas sobre esta temática. Para trabalhos futuros seria interessante a comparação do desempenho deste algoritmo com outros, utilizou-se esse mesmo conjunto de dados.

## REFERÊNCIAS

BOHANEK, M.; RAJKOVIČ, V. DEX: An expert system shell for decision support. **Sistemica**, v. 1, n. 1, p. 145-157, 1990.

BORGES, A. P. **Descoberta de Regras de Condução de Trens de Carga**. 2009. Tese de Doutorado. Pontifícia Universidade Católica do Paraná.

BOUCKAERT, R. R. *et al.* WEKA---Experiences with a Java Open-Source Project. **The Journal of Machine Learning Research**, v. 11, p. 2533-2541, 2010.

CHRISTOPHER, J. J.; NEHEMIAH, H. K.; KANNAN, A. A swarm optimization approach for clinical knowledge mining. **Computer methods and programs in biomedicine**, v. 121, n. 3, p. 137-148, 2015.

COHEN, W. W. Fast effective rule induction. In: **Machine learning proceedings 1995**. Morgan Kaufmann, 1995. p. 115-123.

D'OTTAVIANO, I. M. L.; FEITOSA, H. A.. **História da lógica e o surgimento das lógicas não clássicas**. Coleção História da Matemática para Professores, SBHM/UNESP, v. 1, p. 01-66, 2009.

DOUBRAVSKY, K.; DOHNAL, M. Reconciliation of decision-making heuristics based on decision trees topologies and incomplete fuzzy probabilities sets. **Plos One**, v. 10, n. 7, p. e0131590, 2015.

EVALE, D. S. *et al.* Prediction model for students' performance in Java programming with coursecontent recommendation system. In: **Proceedings of 2016 Universal Technology Management Conference (UTMC)**. Minnesota, United States of America. 2016.

FAIZAL, M. A. *et al.* An analysis of system calls using J48 and jRip for malware detection. **Journal of Theoretical & Applied Information Technology**, v. 96, n. 13, 2018.

FORBELLONE, A. L. V.; EBERSPÄCHER, H. F. **Lógica de programação: a construção de algoritmos e estruturas de dados**. Editora Pearson Prentice Hall. 2021.

GOVADA, A. *et al.* Distributed multi-class rule-based classification using RIPPER. In: **2016 IEEE International Conference on Computer and Information Technology (CIT)**. IEEE, 2016. p. 303-309.

GUVENIR, H. A. *et al.* A supervised machine learning algorithm for arrhythmia analysis. In: **Computers in Cardiology 1997**. IEEE, 1997. p. 433-436.

GÜVENIR, H. A.; DEMİRÖZ, G.; ILTER, N. Learning differential diagnosis of erythematousquamous diseases using voting feature intervals. **Artificial intelligence in medicine**, v. 13, n. 3, p. 147-165, 1998.

HALL, M. *et al.* The WEKA data mining software: an update. **ACM SIGKDD explorations newsletter**, v. 11, n. 1, p. 10-18, 2009.

KSHIRSAGAR, V.; JOSHI, M. Enhancing Intrusion Detection System by Reducing the False Positives through Application of Various Data Mining Techniques. **International Journal of Computer Science and Information Security**, v. 14, n. 2, p. 76, 2016.

KUMAR, V.; VELIDE, L.. A data mining approach for prediction and treatment of diabetes disease. **Int J Sci Invent Today**, v. 3, p. 73-9, 2014.

LAKMALI, K. B. N.; HADDELA, P. S. Effectiveness of rule-based classifiers in Sinhala text categorization. In: **2017 National Information Technology Conference (NITC)**. IEEE, 2017. p. 153-158.

LINS, E. M. V.; WEBBER, C. G.. Integração entre Bases de Casos e Machine Learning: Uma proposta de Help Desk Inteligente. **Scientia cum Industria**, v. 7, n. 2, p. 148-155, 2019.

MANIMARAN, R.; VANITHA, D. M. Novel approach to prediction of diabetes using classification mining algorithm. **International Journal of Innovative Research in Science, Engineering and Technology**, v. 6, n. 7, p. 14481-14487, 2017.

NAKAI, K.; KANEHISA, M.. A knowledge base for predicting protein localization sites in eukaryotic cells. **Genomics**, v. 14, n. 4, p. 897-911, 1992.

PARSANIA, V.; BHALODIYA, N.; JANI, N. N. Applying Naïve bayes, BayesNet, PART, JRip and OneR algorithms on hypothyroid database for comparative analysis. 2014.

QUINLAN, J. R. *et al.* Inductive knowledge acquisition: a case study. In: **Proceedings of the Second Australian Conference on Applications of expert systems**. 1987. p. 137-156.

RAJPUT, An. *et al.* J48 and JRIP rules for e-governance data. **International Journal of Computer Science and Security (IJCSS)**, v. 5, n. 2, p. 201, 2011.

ROMÃO, W.. *Descoberta de conhecimento relevante em banco de dados sobre ciência e tecnologia*, Doutorado em engenharia de produção, Centro Tecnológico da Universidade Federal de Santa Catarina, SC, Brasil, 2002.

SCHLIMMER, J. C.. **Concept acquisition through representational adjustment**. University of California, Irvine, 1987.

SHESKIN, D. J. **Hbook of Parametric Nonparametric Statistical Procedures**, Chapman Hall/CRC, Boca Raton, Florida. 2011.

SMITH, J. W. *et al.*. Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. In: **Proceedings of the annual symposium on computer application in medical care**. American Medical Informatics Association, 1988. p. 261.

SOCZEK, F. C.; ORLOVSKI, R.. *Mineração de Dados: Conceitos e aplicação de algoritmos em uma Base de Dados na área da saúde*. 2014.

SOUZA, P. B.. *Uma estratégia baseada em algoritmos de mineração de dados para validar plano de operação de voo a partir de previsões de estados dos satélites do INPE*. **São José dos Campos, SP, Brasil: INPE**, 2011.

TARUN, I. M.; GERARDO, B. D.; TANGUILIG III, B. T.. Generating licensure examination performance models using PART and JRip classifiers: A data mining application in education. **International Journal of Computer and Communication Engineering**, v. 3, n. 3, p. 203, 2014.

WASEEM, S.; SALMAN, A.; MUHAMMAD, Asif Khan.. Feature subset selection using association rule mining and JRip classifier. **International Journal of Physical Sciences**, v. 8, n. 18, p. 885-896, 2013.

WITTEN, I. H.; FRANK, E.. Data mining: practical machine learning tools and techniques with Java implementations. **Acm Sigmod Record**, v. 31, n. 1, p. 76-77, 2002.

WOOLERY, L. *et al.* The use of machine learning program LERS-LB 2.5 in knowledge acquisition for expert system development in nursing. **Computers in nursing**, v. 9, n. 6, p. 227-234, 1991.

ZWITTER, M.; SOKLIC, M. **Breast cancer data**, Institute of Oncology, University Medical Centre Ljubljana, Yugoslavia. 1988.