Uma ferramenta para geração de datasets educacionais no formato Weka

Pedro D. N. Silveira[1], Davidson Cury[1], Crediné Menezes[1], Otávio L. Santos[1]

[1]Departamento de Informática
Universidade Federal do Espírito Santo (UFES) - Vitória, ES - Brasil

{pedro.dns, dedecury, credine, otaviolube}@gmail.com

Abstract. *Models for data mining (DM) have been extensively discussed in recent years. A software, widely used in academia, that supports the MD process is the Weka. This article presents a tool (and its evaluation) that aims to replace the manual work, exercising the process of transforming the data of a database into a dataset in the format required by Weka. The evaluation of the tool, performed by researchers in the area of MD for education, resulted in positive levels of usability and utility perception.*

Resumo. *Modelos para mineração de dados (MD) têm sido amplamente discutidos nos últimos anos. Um software, muito difundido no meio acadêmico, que auxilia o processo de MD é o Weka. Este artigo apresenta uma ferramenta (e sua avaliação) que objetiva substituir o trabalho manual, exercendo o processo de transformação dos dados de um banco de dados, em um dataset no formato exigido pelo Weka. A avaliação da ferramenta, realizada por pesquisadores da área de MD para educação, resultou em níveis positivos de usabilidade e percepção de utilidade.*

1. Introdução

Mineração de dados educacionais (em inglês Educational Data Mining - EDM) é uma tecnologia em expansão com uma grande diversidade de aplicativos disponíveis. Está se tornando um dos principais componentes de sistemas inteligentes de informação para auxílio da tomada de decisão em termos educacionais.

Algumas técnicas de aprendizagem de máquina, amplamente difundidas na literatura, como por exemplo regressão logística [Hosmer Jr et al. 2013] são usadas para prever situações acadêmicas como desempenho [Adeodato et al. 2014], abandono [Calixto et al. 2017], quantidade de inscrições em cursos [Haddawy et al. 2007] dentre outros.

Existem várias ferramentas que fazem a aplicação das técnicas de aprendizagem de máquina e consequentemente, mineração de dados, dentre as quais se destaca o sistema *Waikato Environment for Knowledge Analysis* (Weka) [Garner et al. 1995]. Normalmente, essas ferramentas exigem uma configuração específica para o conjunto de dados com o qual irão trabalhar, que pode ser um simples arquivo .csv (Comma-Separated Values) ou um arquivo com formato bem mais elaborado, específico da ferramenta.

Nosso grupo de pesquisa, atualmente tem trabalhado em larga escala com as ecologias cognitivas que em sua essência primam fortemente pela interação

DOI: 10.5753/cbie.sbie.2018.1343

de seus participantes (estudantes e professores) entre si e com o ambiente [Dukas 1998].Temos registrado essas interações por meio do armazenamento dos dados de acesso e da utilização das diversas ferramentas presentes nos ambientes de ensino-aprendizagem com os quais trabalhamos, em especial o Moodle. A partir disso, tem-nos sido possível proceder com a análise de aprendizagem (*learning analytics*).

Durante a nossa pesquisa, utilizando o Weka, vivenciamos uma dificuldade que é responsável por tomar uma considerável parte do tempo de pesquisa, que é a obtenção de um arquivo no formato que a ferramenta exige a partir dos dados existentes em nossos bancos de dados. Isto é, precisávamos automatizar a geração do arquivo de dados do Weka a partir de uma consulta SQL (Structured Query Language).

Com essa percepção, objetivamos o desenvolvimento de uma ferramenta que permita a conexão ou importação de um banco de dados SQL e forneça uma interface para realização de busca de informação, de forma que o resultado da consulta seja um arquivo de texto contendo um *dataset* no padrão exigido pela ferramenta Weka, colocando fim a manualidade deste trabalho. A ferramenta já se encontra em uso, com resultados promissores.

Este documento está organizado da seguinte forma: a Seção 2 apresenta o referencial teórico e os trabalhos correlatos. Na Seção 3 são destacados os resultados obtidos no estudo e a metodologia de desenvolvimento da aplicação. Finalmente, a Seção 4 apresenta as considerações finais e trabalhos futuros.

2. Referencial teórico e trabalhos correlatos

Nesta seção serão abordados conceitos essenciais ao entendimento da proposta descrita neste documento, destacando-se ecologias cognitivas, mineração de dados educacionais e *Learning Analytics* (LA), a ferramenta Weka, formatos de arquivo de mineração de dados e também os trabalhos correlatos.

O conceito de ecologia da mente [Bateson 1986] advém de uma teoria epistêmica que é baseada na concepção das interações de indivíduos e o ambiente no qual estão inseridos, que podem ser inclusive, virtuais. [Smart et al. 2017], acreditam que os ambientes cooperativos online, seja para ensino-aprendizagem ou sociais, estão enquadrados em um contexto onde as diversas formas de cognição podem ser exploradas ou potencializadas. Nesse contexto estão as chamadas ecologias cognitivas, que podem ser definidas como "os contextos multidimensionais em que lembramos, sentimos, pensamos, comunicamos, imaginamos e agimos, com frequência em colaboração, na vida e em rica e contínua interação com nossos ambientes". [Tribble and Sutton 2011]

A partir dessa discussão e da epistemologia genética [Piaget and Del Val 1970] percebe-se uma forte ligação das interações com o processo de aprendizagem. O processo de interação entre indivíduos possibilita trocar pontos de vista, conhecer e refletir sobre diferentes questionamentos e seu próprio pensar, e então ampliar com autonomia uma tomada de consciência para buscar novos rumos [Tijiboy et al. 1999]. Havendo um registro dessas interações, torna-se possível a geração de *datasets* para de mineração de dados.

Mineração de dados, está diretamente ligada à Descoberta de Conhecimentos em Bancos de Dados, ou KDD (do inglês, *Knowledge Discovery in Databases*). Para [Zyt et al. 2002], KDD é uma análise e exploração de grande volume de dados para a descoberta automática de conhecimentos e refere-se a todo o processo de descoberta de conhecimento útil em dados. A mineração de dados, que é uma etapa deste processo, é o uso de algoritmos de aprendizado de máquinas para procurar o conhecimento valioso.

A mineração de dados aplicada a educação é definida como a área de pesquisa que objetiva o desenvolvimento e aplicação de métodos para explorar conjuntos de dados coletados em ambientes educacionais. Com isso torna-se viável compreender de forma mais eficaz e adequada os alunos, como eles aprendem, o papel do contexto na qual a aprendizagem ocorre, além de outros fatores que influenciam a aprendizagem. [Baker et al. 2011]

A larga adoção de tecnologias educacionais proporcionou uma nova oportunidade de obter descobertas sobre o aprendizado dos alunos. Assim como na maioria dos sistemas de TI, as interações do aluno com suas atividades de aprendizado online são capturadas e armazenadas. Esses traços digitais (dados de registro ou dados de rastreio) podem ser extraídos e analisados para identificar padrões de comportamento de aprendizagem. Esse processo foi descrito como *learning analitics*. [Gašević et al. 2015]

Recentemente, LA atraiu a atenção de acadêmicos, pesquisadores e administradores como suporte tecnológico para a apoiar a aprendizagem no contexto das ecologias cognitivas. Esse interesse é motivado pela necessidade aprimorar o ensino, dadas as atuais possibilidades tecnológicas capazes de promoverem a integração e socialização de todo conhecimento produzido por um grupo ou mesmo um único indivíduo. Organizações como *Society for Learning Analytics Research* e *International Educational Data Mining Society* foram criadas para promover uma comunidade de pesquisa em torno do papel da análise de dados na educação para esse fim.

No contexto deste trabalho, usamos como ferramenta para mineração de dados o software Weka [Garner et al. 1995]. O Weka foi projetado para oferecer uma variedade de técnicas ou esquemas de *machine learning* sob uma interface comum, para que possam ser facilmente aplicados a um conjunto de dados de maneira consistente. Atualmente, o Weka alcançou aceitação generalizada nos círculos acadêmicos e empresariais, e se tornou uma ferramenta amplamente usada para pesquisa de mineração de dados. [Hall et al. 2009]

O sistema Weka usa um formato de arquivo específico para trabalhar seus conjuntos de dados, independentemente do esquema de aprendizado de máquina que possa ser usado. Esse formato de arquivo, o Formato de Arquivo de Relação de Atributos, do inglês *Attribute-Relation File Format* (ARFF), define um conjunto de dados em termos de uma tabela composta de atributos. Informações sobre os nomes da tabela e os tipos de dados dos atributos são armazenados no cabeçalho ARFF, e as instâncias são representadas como linhas de dados no corpo do arquivo ARFF. [Garner et al. 1995]

```
@relation golf

@attribute outlook { sunny, overcast, rain}
@attribute temperature real [0.0,100]
@attribute humidity real
@attribute windy { true, false}
@attribute class { Play, 'Dont Play' }

@data
% 14 instances follow
sunny, 85, 85, false, 'Dont Play'
sunny, 80, 90, true, 'Dont Play'
overcast, 83, 78, false, Play
rain, 70, 96, false, Play
rain, 68, 80, false, Play
rain, 65, 70, true, 'Dont Play'
overcast, 64, 65, true, Play
sunny, 72, 95, false, 'Dont Play'
sunny, 69, 70, false, Play
rain, 75, 80, false, Play
sunny, 75, 70, true, Play
overcast, 72, 90, true, Play
overcast, 81, 75, false, Play
rain, 71, 80, true, 'Dont Play'
```

Figura 1. Exemplo de arquivo .arff [Garner et al. 1995]

Atualmente, é permitido aos atributos aceitar três tipos diferentes: inteiros, numéricos (ponto flutuante) e enumerações. Com os atributos numéricos, um intervalo opcional pode ser especificado e, caso necessário, os atributos booleanos podem ser tratados como uma enumeração com dois valores. Um exemplo de arquivo ARFF é mostrado na Figura 1.

2.1. Trabalhos correlatos

Como trabalhos correlatos, [Omari et al. 2008] cita em seu artigo algumas ferramentas geradoras de *datasets* como, por exemplo *E-commerce Generator* [Groblschegg 2003] e *ARtool Generator* [Cristofor 2008] que produzem conjuntos de dados para uma cesta de mercado de comércio eletrônico. No entanto essas ferramentas são para geração de conjunto de dados e não para transformação dos mesmos em um formato específico.

Algumas ferramentas para geração de *datasets* no formato Weka estão disponíveis, como CSV2ARFF[1] e a classe Java (do próprio Weka) AttTest[2]. No entanto, a primeira não se conecta a um sistema gerenciador de banco de dados, ou seja, não existe uma consulta SQL atrelada à geração do arquivo ARFF, mas uma conversão direta de um arquivo no formato CSV para o formato weka. A segunda pode até receber como entrada uma consulta SQL, mas exige um software em desenvolvimento na linguagem java e consequente conhecimento prévio em programação para utilização e implantação.

A tabela 1 apresenta uma comparação do Gerador de Datasets Weka (GDW) que é a ferramenta que estamos propondo aqui com as ferramentas correlatas relacionadas no texto, sob o ponto de vista de entradas e saídas desejadas.

Até o término da escrita deste documento, não foram encontradas propostas ou ferramentas que executem a geração de arquivos ”.arff”a partir de consultas SQL em um banco de dados. A chave de busca aplicada nas bases ACM

[1]Disponível em http://ikuz.eu/csv2arff/
[2]Disponível em https://waikato.github.io/weka-wiki/creating_arff_file/

Tabela 1. Comparação do GDW com as ferramentas correlatas

Ferramenta	Saída: ARFF	Entrada: SQL	Sem programação
E-commerce Generator		X	X
ARtool Generator		X	X
CSV2ARFF	X		
AttTest	X	X	
GDW	X	X	X

Digital Library, IEEE Xplore e Science Direct foi: *(".arff file"OR "weka file") AND "SQL Query"*.

3. Metodologia

Para a produção deste documento foram realizadas algumas etapas. Primeiro um de levantamento bibliográfico tipo exploratório foi executado objetivando alcançar os conhecimentos teóricos e técnicos sobre o formato de arquivo de dados Weka e bibliotecas de manipulação de grande massa de dados. Posteriormente realizamos a etapa de estudo de trabalhos correlatos. Em seguida buscamos estabelecer a motivação e objetivos citados anteriormente, assim como a solução proposta conceitualmente. Com base nessas informações, demos início ao processo de desenvolvimento da ferramenta que chamamos de "Gerador de Dataset Weka"(GDW).

Realizamos o levantamento de requisitos e modelagem do sistema, posteriormente a implementação e por fim os testes. O gerador de arquivos .arff foi construído seguindo o modelo de processo de desenvolvimento de software iterativo e incremental. A parte administrativa do site, isto é toda a parte de interação com o usuário, foi programada com PHP, Javascript, CSS e HTML. As funções de transformação dos dados (nativos no banco de dados) no arquivo .arff, bem como a inteligência artificial que gera a consulta SQL a partir da seleção efetuada pelos usuários, foram implementadas em python.

Após o desenvolvimento da ferramenta, a disponibilizamos online e foram realizados testes de avaliação com um grupo de 10 pesquisadores da área de informática na educação (mais especificamente trabalhando com ecologias cognitivas) que são usuários da ferramenta Weka e atualmente utilizam mineração de dados educacionais como um instrumento de pesquisa. Cada um deles respondeu um formulário de avaliação com cinco afirmações sobre a percepção de usabilidade e utilidade da ferramenta.

O usuário deveria corresponder às afirmações da Tabela 2 com: "Discordo totalmente", "Discordo parcialmente", "Indiferente", "Concordo parcialmente", "Concordo totalmente", obedecendo uma escala Likert [Allen and Seaman 2007], que estabelece um peso de 1 a 5 respectivamente, para cada afirmação.

4. Resultados e Discussão

Na primeira parte desta seção, apresentaremos a ferramenta Gerador de Datasets Weka e seu funcionamento. Na segunda parte mostraremos e discutiremos os

Tabela 2. Pesquisa com usuários

	Pergunta
A1.	O GDW é intuitivo e fácil de usar
A2.	O GDW não precisa de modificações
A3.	O GDW é útil em mineração de dados educacionais
A4.	O GDW vai me ajudar em algum momento da minha pesquisa
A5.	Não conheço nenhuma ferramenta similar ao GDW em seu propósito

resultados da avaliação da ferramenta por parte daqueles que a usaram.

4.1. A ferramenta Gerador de Datasets Weka (GWD)

Após o login no sistema, é fornecido ao usuário duas opções. Na primeira, a ferramenta GDW permite que seja feita uma conexão a um banco de dados existente (tela do sistema demonstrada na Figura 2-B), e na segunda, que seja importado um arquivo .sql contendo um *dump* de um banco de dados (MySQL) externo para o banco de dados disponível no servidor onde o GDW está disponível (Figura 2-C). A partir da conexão ou da importação do banco de dados, o usuário é redirecionado para o executor de consultas SQL do GDW (Figura 3).

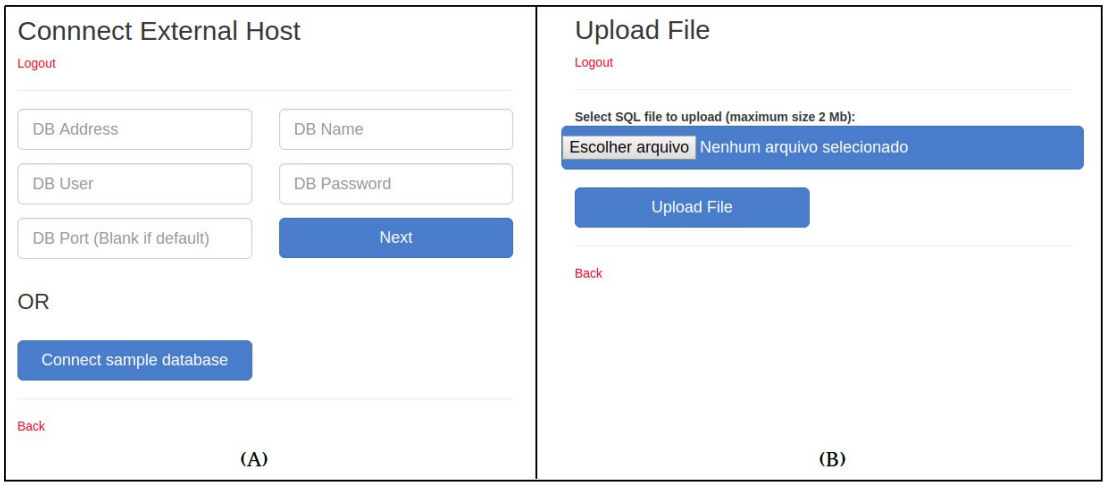


Figura 2. Modos de operação GDW.

Ao carregar a tela da Figura 3 o sistema mostra todas as tabelas (e seus respectivos atributos) disponíveis no banco de dados (importado ou conectado). Nas tabelas apresentadas também é informado ao usuário quais atributos da tabela são as chaves primárias ou estrangeiras.

As tabelas demonstradas na Figura 3 estão em um banco de dados, no servidor do GDW, disponível para testes e podem ser acessadas clicando em "*Connect sample database*"na tela da Figura 2-A.

Figura 3. Manipulador SQL do GDW

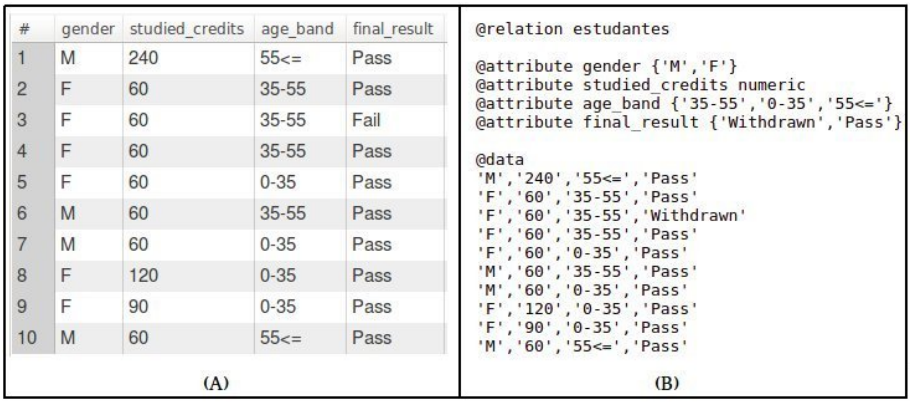
O usuário deve então fornecer o nome do arquivo .arff a ser gerado e providenciar a consulta SQL. A consulta pode ser inserida manualmente ou gerada automaticamente a partir da seleção dos atributos (clique) nas próprias tabelas, como observado na Figura 3 em alguns campos da tabela "*student*".

Para a geração da consulta automática, foi necessária a implementação de um algoritmo inteligente, que identifica como diferentes tabelas se tornam alcançáveis a partir da seleção de seus atributos. O algoritmo faz as verificações pelo caminhamento entre as chaves primárias e estrangeiras nas permutações de todas as tabelas da base de dados, e no fim gera uma consulta sem filtros, que podem ser inseridos manualmente se for o caso.

O usuário deve ter em mente que o resultado de uma consulta aplicada em um banco de dados qualquer, nada mais é que uma tabela de dados (*result set*). A ferramenta GDW se encarrega de transformar esse *result set* em um arquivo que o Weka será capaz de interpretar. A Figura 4-A mostra o resultado de uma consulta SQL aplicada diretamente no banco de dados, enquanto na 4-B podemos ver o conteúdo do arquivo gerado com o processamento interno do GDW com os mesmos parâmetros da consulta (que estão presentes no formulário da tela da Figura 3).

4.2. Avaliação da ferramenta

A Tabela 3, contem o resultado da pesquisa realizada com os usuários do GDW já sumarizado em média (± desvio padrão) de todas as respostas. Cada asser-



#	gender	studied_credits	age_band	final_result
1	M	240	55<=	Pass
2	F	60	35-55	Pass
3	F	60	35-55	Fail
4	F	60	35-55	Pass
5	F	60	0-35	Pass
6	M	60	35-55	Pass
7	M	60	0-35	Pass
8	F	120	0-35	Pass
9	F	90	0-35	Pass
10	M	60	55<=	Pass

(A)

```
@relation estudantes

@attribute gender {'M','F'}
@attribute studied_credits numeric
@attribute age_band {'35-55','0-35','55<='}
@attribute final_result {'Withdrawn','Pass'}

@data
'M','240','55<=','Pass'
'F','60','35-55','Pass'
'F','60','35-55','Withdrawn'
'F','60','35-55','Pass'
'F','60','0-35','Pass'
'M','60','35-55','Pass'
'M','60','0-35','Pass'
'F','120','0-35','Pass'
'F','90','0-35','Pass'
'M','60','55<=','Pass'
```

(B)

Figura 4. Resultado da consulta SQL x Arquivo .arff gerado com GDW

tiva (de A1 até A5) poderia receber 5 como valor máximo de avaliação e 1 como valor mínimo em termos de concordância com a afirmação (como explicado na metodologia).

Tabela 3. Sumarização da avaliação da ferramenta GDW

	A1	**A2**	**A3**	**A4**	**A5**
Avaliação Média	3,9	2,2	4,8	4,0	4,6
Desvio Padrão	$\pm 1,1$	$\pm 0,9$	$\pm 0,4$	$\pm 1,2$	$\pm 0,7$

O objetivo das assertivas A1 e A2, tem caráter de avaliação de usabilidade do sistema. Com o resultado obtido para A1, obtivemos indícios que a interface do sistema é adequada, permitindo ele seja usado de modo direto e instantâneo, naturalmente. Já com o resultado para a assertiva A2, percebemos a necessidade de melhoria nas funcionalidades do site, que inclusive foram sugeridas pelos próprios usuários (no questionário, o usuário pôde apresentar sugestões de modificação, em forma textual, na usabilidade do GDW).

O objetivo das assertivas A3 e A4 tem caráter de avaliação da utilidade da ferramenta. Com os resultados obtidos para A3 e A4, alcançamos indícios de que a ferramenta contribui com pesquisadores que atuam com mineração de dados educacionais, diminuindo seu esforço na geração de *datasets* no formato Weka a partir de seus bancos de dados. Já com o resultado obtido para A5, percebemos a necessidade de existência do GDW em seu propósito.

Na escala que adotamos, uma média que se aproxima de 3 (ponto médio) nos indica uma neutralidade na resposta, que acreditamos que deve ser considerada, tendo em vista que [Worcester and Burns 1975] concluíram em seus estudos que uma escala de quatro pontos sem um ponto médio parece empurrar mais respondentes para o final positivo da escala, criando um enviesamento na pesquisa.

5. Conclusão

Neste artigo apresentamos uma ferramenta para geração de *datasets* no formato Weka. A necessidade de implementação do sistema surgiu a partir da custosa

e constante transformação manual (no formato .arff) de resultados de consultas SQL em dados gerados no projeto de ecologias cognitivas a partir das interações de indivíduos entre si e com os ambientes virtuais para ensino-aprendizagem.

Nossa contribuição está na disponibilização da ferramenta[3] para pesquisadores que trabalhem com mineração de dados e utilizem o software Weka. Os resultados da avaliação, nos mostram preliminarmente que estamos indo na direção correta mesmo com as lacunas existentes que ainda podem ser exploradas em trabalhos futuros. Cabe ainda enfatizar nosso interesse particular nessa ferramenta pelo no fato de estarmos desenvolvendo, ainda em fase de especificação de requisitos, um meta-ambiente para dar suporte ao projeto de ferramentas operantes em ecologias cognitivas.

Em uma segunda versão da ferramenta, pretendemos uma atualização para permitir a conexão com outros sistemas gerenciadores de banco de dados e não apenas o MySQL. Também temos a intenção de acoplar o GDW ao Weka (que é software livre com código fonte aberto) disponibilizando uma ferramenta integrada.

Referências

Adeodato, P. J., Santos Filho, M. M., and Rodrigues, R. L. (2014). Predição de desempenho de escolas privadas usando o enem como indicador de qualidade escolar. In *Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)*, volume 25, page 891.

Allen, I. E. and Seaman, C. A. (2007). Likert scales and data analyses. *Quality progress*, 40(7):64.

Baker, R., Isotani, S., and Carvalho, A. (2011). Mineraçao de dados educacionais: Oportunidades para o brasil. *Brazilian Journal of Computers in Education*, 19(02):03.

Bateson, G. (1986). *Mente e natureza: a unidade necessária*. F. Alves.

Calixto, K., Segundo, C., and de Gusmão, R. P. (2017). Mineração de dados aplicada a educação: um estudo comparativo acerca das características que influenciam a evasão escolar. In *Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)*, volume 28, page 1447.

Cristofor, L. (2008). Artool project. university of massachusetts, boston.

Dukas, R. (1998). *Cognitive ecology: the evolutionary ecology of information processing and decision making*. University of Chicago Press.

Garner, S. R. et al. (1995). Weka: The waikato environment for knowledge analysis. In *Proceedings of the New Zealand computer science research students conference*, pages 57–64. Citeseer.

Gašević, D., Dawson, S., and Siemens, G. (2015). Let's not forget: Learning analytics are about learning. *TechTrends*, 59(1):64–71.

[3]Temporariamente disponível em http://200.137.66.25/GDW/

Groblschegg, M. (2003). Developing a testdata generator for market basket analysis for e-commerce applications. *Vienna University of Economics and Business Administration*.

Haddawy, P. et al. (2007). Deriving financial aid optimization models from admissions data. In *Frontiers In Education Conference-Global Engineering: Knowledge Without Borders, Opportunities Without Passports, 2007. FIE'07. 37th Annual*, pages F2A–7. IEEE.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The weka data mining software: an update. *ACM SIGKDD explorations newsletter*, 11(1):10–18.

Hosmer Jr, D. W., Lemeshow, S., and Sturdivant, R. X. (2013). *Applied logistic regression*, volume 398. John Wiley & Sons.

Omari, A., Langer, R., and Conrad, S. (2008). Tartool: A temporal dataset generator for market basket analysis. In *International Conference on Advanced Data Mining and Applications*, pages 400–410. Springer.

Piaget, J. and Del Val, J. A. (1970). *La epistemología genética*. A. Redondo.

Smart, P., Heersmink, R., and Clowes, R. W. (2017). The cognitive ecology of the internet. In *Cognition beyond the brain*, pages 251–282. Springer.

Tijiboy, A. V., Maçada, D. L., Santarosa, L. M. C., and Fagundes, L. d. C. (1999). Aprendizagem cooperativa em ambientes telemáticos. *Informática na Educação: teoria & prática. Porto Alegre. Vol. 1, n. 2 (abr. 1999), p. 19-28*.

Tribble, E. and Sutton, J. (2011). Cognitive ecology as a framework for shakespearean studies. *Shakespeare Studies*, 39:94.

Worcester, R. M. and Burns, T. R. (1975). Statistical examination of relative precision of verbal scales. *Journal of the Market Research Society*, 17(3):181–197.

Zyt, J., Klosgen, W., and Zytkow, J. (2002). *Handbook of data mining and knowledge discovery*. Oxford university press.