Mineração dos Dados do Enade: Avaliação do questionário do estudante das Instituições do Norte do Brasil

Alexandre Moraes Matos[1], Lucas Ribeiro Reis de Sousa[1], Ismael Pontes Torres Júnior[1], Renato Marinho Alves[1], Heloise Acco Tives Leão[2], Fabiano Fagundes[1]

[1]Departamento de Computação – CEULP/ULBRA – Palmas/TO

[2]Instituto de Ciências Exatas - Universidade de Brasília - (UnB)
Brasília - DF - Brasil

{alexandremt03, lucasguitar45, ismaelpontesjr, renato.mar.alves, heloise.acco, thilfa}@gmail.com

Resumo. *O Instituto Nacional de Estudos e Pesquisas Educacionais (INEP) disponibiliza dados do Exame Nacional de Desempenho de Estudantes (ENADE) para as Instituições de Ensino Superior (IES) do Brasil. Esses dados são uma fonte muito importante de informação para auxiliar na melhora da qualidade do ensino superior oferecido por essas IES. Para isso precisam ser aplicadas técnicas de mineração de dados para o alcance de padrões do processo de aprendizagem e com isso podermos alcançar uma melhora no desempenho acadêmico dos estudantes em diferentes cursos. Este trabalho apresenta as etapas de mineração dos dados fornecidos pelo INEP, dos IES do Norte do Brasil, com o objetivo de identificar padrões e propor melhorias que visam a qualidade do ensino e satisfação dos discentes.*

1. Introdução

O Sistema Nacional de Avaliação da Educação Superior (Sinaes) faz uso do Exame Nacional do Desempenho dos Estudantes (ENADE) para avaliar os cursos do ensino superior. Ele é realizado anualmente desde 2004 pelo Instituto Nacional de Estudantes e Pesquisa Educacional Anísio Teixeira (INEP).

A partir de uma amostra selecionada de estudantes do primeiro e do último ano dos cursos de graduação das instituições de Ensino Superior do Brasil o exame é realizado. Além de avaliar a qualidade dos cursos de formação superior, também tem como objetivo fazer uma classificação única para os cursos de graduação do Brasil.

As etapas que compõem o Enade são o Questionário do Estudante, a Prova e o Questionário da Coordenação do Curso. Este trabalho irá realizar e apresentar a mineração dos dados coletados, organizados e disponibilizados pelo Inep acerca do questionário do Estudante, com o objetivo de identificar as queixas mais relevantes dos discentes e a partir disso propor alternativas de melhoria no processo de ensino.

Para isso, foi selecionado como período de análise os anos de 2014 a 2016 e os dados de todas as Instituições da Região Norte do Brasil que participaram do Enade, sendo no ano de 2014, 109 instituições; no ano de 2015, 116 instituições e no ano de 2016, 88 instituições; totalizando 137 instituições diferentes no decorrer de todo o período avaliado neste projeto de mineração.

Este trabalho está organizado da seguinte forma: na seção 2 será descrito o referencial teórico necessário para embasar a realização do trabalho. Na seção 3 é apresentada a metodologia utilizada no desenvolvimento do trabalho. Na seção 4 são descritos os resultados da mineração dos microdados do ENADE. Por fim, a seção 5 traz as conclusões obtidas durante o desenvolvimento deste trabalho.

ISSN: 2447-0767

2. Referencial Teórico

A Mineração de Dados é uma etapa de extrema importância para que haja o descobrimento de informações, tendo em vista que há uma grande quantidade de informações potenciais que podem ser obtidas a partir de análises mais profundas dos dados. Da Silva (2016, p. 7) conceitua a mineração de dados como "um esforço para descoberta de padrões em bases de dados". Observando este conceito, percebe-se que pela descoberta destes padrões é possível a obtenção informações que proporcionem auxílio em tomadas de decisões de negócios ou sua utilização para análise científica.

A utilização de Mineração de Dados no âmbito empresarial e científico vem obtendo grande foco, devido à possibilidade de obtenção de novas informações que possibilitem o desenvolvimento de empresas e/ou projetos. De acordo com Cortês (2002) a mineração de dados é classificada com uma combinação entre pesquisas em estatística, inteligência artificial e bancos de dados, que vem emergindo como uma área de grande importância que destaca-se pelo surgimento de diversos congressos científicos e produtos comerciais.

De acordo com Tan (2009, p.3) a Mineração de Dados se fundamenta na utilização de meios automáticos para a busca e descoberta de padrões em grandes bases de dados. Diante disso são propostos algoritmos computacionais que implementam lógicas de extração e análise de informações.

Os algoritmos de DM em geral procedem sua execução a partir de uma entrada, sendo em geral dados formatados e padronizados, e a partir destes dados e da execução dos algoritmos, resultados são apresentados. A DM trata-se da aplicação de técnicas, implementadas através de algoritmos que são capazes de receber fatos ocorridos como entrada e resultar, como saída, um padrão comportamental, que pode ser apresentado, como exemplo, em regra de associação, função de mapeamento ou modelagem de um perfil [DA SILVA; PERES; BOSCARIOLI, 2016, p. 7].

Os algoritmos implementados para Mineração de Dados geralmente utilizam conceitos de Inteligência Artificial (IA) e Aprendizado de Máquina. Camilo e Silva afirmam que (2009, p.10) os métodos de DM são divididos em aprendizado supervisionado (preditivo) e não-supervisionado (descritivo). De modo que o aprendizado supervisionado exige o acompanhamento constante do analista de dados no processo de mineração, enquanto o processo não-supervisionado propõe que são necessários apenas os dados de entrada para se gerar uma saída.

Dentre os diversos algoritmos de Mineração de Dados e seus usos, o algoritmo de associação Apriori que foi proposto por AGRAWAL [Agrawal et al., 1994] é o mais utilizado atualmente para descobrir regras de associação [LibrelottoandMozzaquatro, 2013], que são consideradas apropriadas para analisar dados educacionais onde pretende-se identificar padrões do processo de aprendizagem e melhorar o desempenho acadêmico de estudantes em diferentes cursos [Mobasher et al., 2017].

O algoritmo Apriori executa múltiplas passagens sobre o banco de dados de transações, e é capaz de trabalhar com um número grande de atributos, obtendo como resultado, várias alternativas combinatórias entre eles, a partir da realização de buscas sucessivas em toda a base de dados e, apesar disso, os autores apontam o ótimo desempenho em termos de processamento desse algoritmo.

Quando se deseja mostrar a frequência e a identificação de padrões em conjuntos de dados, o algoritmo de Apriori é usual, já que com sua aplicação são testados muitas vezes os atributos em busca de regras expandidas que representem padrões de sua população (Jang et al., 2015).

No uso do algoritmo de Apriori, as medidas que influenciam a descoberta das regras são: suporte que e porcentagem de casos em que contém tanto A e B; confiança que é a porcentagem de casos contendo A que também contém B;elift que é a taxa de confiança com a porcentagem de casos contendo B [KalgotraandSharda, 2016].

De maneira a otimizar o processo mineração de dados, existem modelos de referência que propõem estruturar uma série de passos a serem seguidos para a obtenção de um melhor e mais rápido resultado, como o CRISP-DM (*Cross Industry Standard Process for Data Mining*/ Processo Padrão de Vários Segmentos de Mercados para Mineração de Dados). Este modelo de referência consiste em um conjunto de etapas que buscam aumentar a taxa de sucesso de processos de Data Mining. Segundo a IBM (2017) "CRISP-DM é uma forma comprovada pelo mercado para orientar seus esforços de mineração de dados".

O modelo CRISP-DM é trabalhado de maneira cíclica em torno de seis fases. Chapman et. al. (2005, p.6) descrevem a metodologia do modelo como um processo hierárquico, que compõe tarefas divididas em quatro níveis de abstração: fase, tarefa genérica, tarefa especializada e processo de instância. Diante disso a Figura 1 apresenta, em formato gráfico, as fases do CRISP-DM.

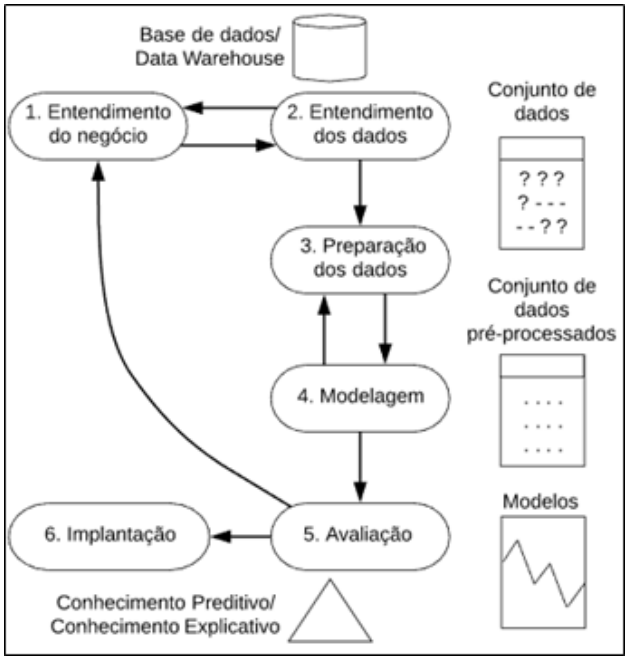

Figura 1 - Modelo de processos do CRISP-DM. Fonte: Adaptado de Moro (2011)

A Figura 1 proposta por Moro (2011) demonstra como é a estruturação das fases do CRISP-DM, sendo que cada uma é essencial para o correto funcionamento do modelo. A seguir essas fases são brevemente descritas:

- **Entendimento do negócio:** Etapa focada no entendimento do objetivo a ser atingido ao se usar a Mineração de Dados. Sendo assim uma etapa fundamental para o desenvolvimento das demais etapas.
- **Entendimento dos dados:** Consiste em compreender os dados que estão sendo analisados de forma a identificar o conjunto de dados relevante à proposta. De acordo com Camilo e Silva (2009) "As fontes fornecedoras dos dados podem vir de diversos locais e possuírem diversos formatos"..
- **Preparação dos dados:** Esta etapa consiste na formatação e transformação dos dados de modo a padronizá-los. Segundo Olson e Delen (2008) o propósito da preparação dos dados é limpar os dados selecionados de modo a obter-se melhor

ISSN: 2447-0767

qualidade, tendo em vista que alguns dos dados selecionados podem seguir diferentes padrões por conta de serem coletados de diferentes fontes.

- **Modelagem:** Na etapa de modelagem são aplicados os algoritmos de mineração de dados de modo que gerar os resultados esperados
- **Avaliação:** "Considerada uma fase crítica do processo de mineração, nesta etapa é necessária a participação de especialistas nos dados, conhecedores do negócio e tomadores de decisão." (CAMILO; SILVA, 2009). Para apoiar nessa etapa são utilizados gráficos para analisar e visualizar os resultados obtidos nas etapas anteriores. Para garantir a confiabilidade dos modelos é indicada a realização de testes e validações nos modelos construídos.
- **Implementação:** Nesta etapa os resultados do projeto de DM são apresentados aos envolvidos. "O estudo de *data mining* possui novos conhecimentos descobertos, que necessitam de estar bem atados aos objetivos originais do projeto de *data mining*." (OLSON; DELEN, 2008).

3. Metodologia

Com o objetivo de aplicar técnicas de mineração sobre os microdados do Questionário dos Estudantes respondidos pelos discentes no final das provas ENADE dos anos de 2014 a 2016, e a partir disso extrair informações relevantes sobre a satisfação desses discentes e identificar queixas dos estudantes ou pontos melhorias a serem analisadas pelas IES do Norte do Brasil, este trabalho foi contextualizado e estruturado.

A metodologia envolve a revisão de literatura que guiou a escolha da técnica de mineração a ser aplicada, a escolha do algoritmo a ser utilizado e a identificação do modelo de referência a ser seguido.

Com base em pesquisas comparativas das técnicas de mineração de dados [Jang et al., 2015], [Luna et al., 2017], [woo Kim et al., 2017] e de mineração de dados da área educacional [Mobasher et al., 2017], [OugiaroglouandPaschalis, 2012], [KumarandChadha, 2012], [Hoed, 2017], [Leao, et al, 2018] foi escolhido o algoritmo de Apriori para resolução da tarefa de descoberta de regras de associação neste trabalho.

As demais etapas a serem desenvolvidas no trabalho, consistem em conhecer o ambiente a que os dados se referem, utilização do modelo de referência CRISP-DM [KalgotraandSharda, 2016] para execução das etapas de mineração, execução de testes para avaliar os resultados encontrados, apresentar resultados e conclusões do projeto assim como as sugestões de formas de implantar melhorias para aumentar a satisfação dos acadêmicos.

Os dados a serem utilizados para esta pesquisa foram extraídos dos microdados do ENADE dispostos no website do Inep em arquivos no formato .csv, que contém informações sobre os acadêmicos, suas respostas e os gabaritos respectivos para as provas de cada área de todo o país.

Para que este trabalho fosse desenvolvido foi utilizado o Microsoft Excel como ferramenta para tratamento de dado, tendo em vista que este permite a limpeza e tratamento dos dados de forma dinâmica e simples. Para a análise foram utilizadas as ferramentas **R Studio e Weka**, juntamente a **linguagem de programação R**, onde foram utilizados os dados em um arquivo no formato .csv fornecido pelo Inep.

4. Estudo de caso

Neste estudo de caso foram identificadas **20 associações** para o ano de 2014, **72 associações** para 2015 e **84 associações** para 2016 pelo algoritmo Apriori. Com apoio da ferramenta

RStudio, foram gerados gráficos que permitiram a análise e melhor entendimento dos padrões descobertos.

As regras encontradas foram configuradas com suporte e confiança acima de 75%, para evitar grande quantidade de regras e possibilitar uma melhor compreensão dos padrões realmente relevantes.

O resultado do ano de 2014 é apresentado na figura 2, onde é possível verificar as associações geradas pelo algoritmo. Percebe-se que que 79% dos estudantes que responderam "letra a" na questão 14 e 17 também responderam "letra a" na questão 3. Verifica-se com isto que 79% dos discentes tem o mesmo perfil, sendo formado por brasileiros, que estudaram todo ensino médio em escola pública e não participaram de programas acadêmicos no exterior.

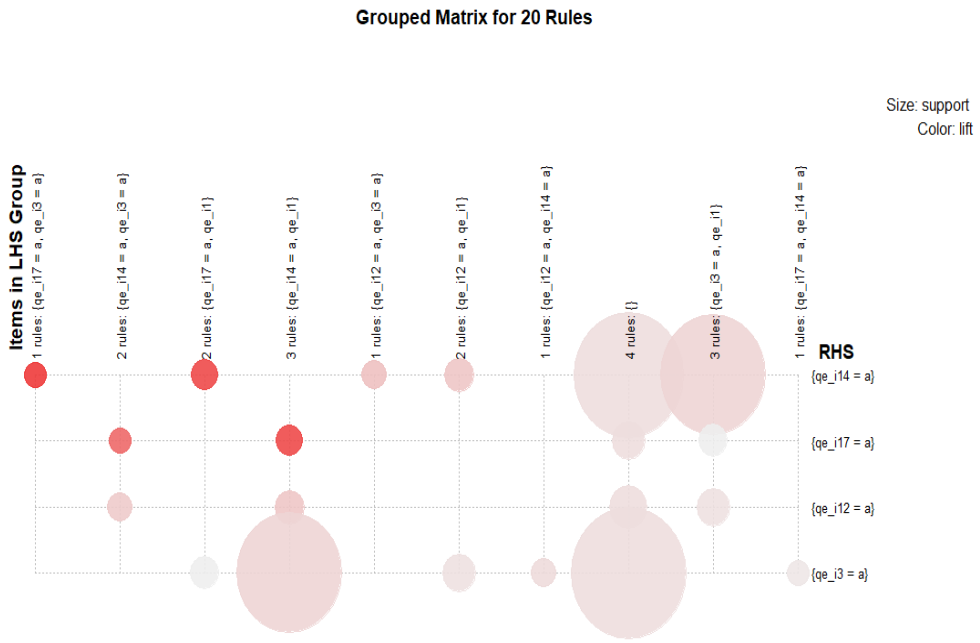

Figura 2. Matriz de visualização para 20 regras no ano de 2014.

Aplicando a mesma análise para o ano de 2015, foram identificados padrões como por exemplo: 77% dos estudantes que responderam que não possuem nenhum auxílio na questão 12, que nunca participaram de programas de ensino na questão 14 e que alegaram ser brasileiros na questão 3, concluíram o ensino médio de forma tradicional. A figura 3, apresenta o gráfico que demonstra o nível de associação a partir das 72 regras geradas no ano de 2015.

ISSN: 2447-0767

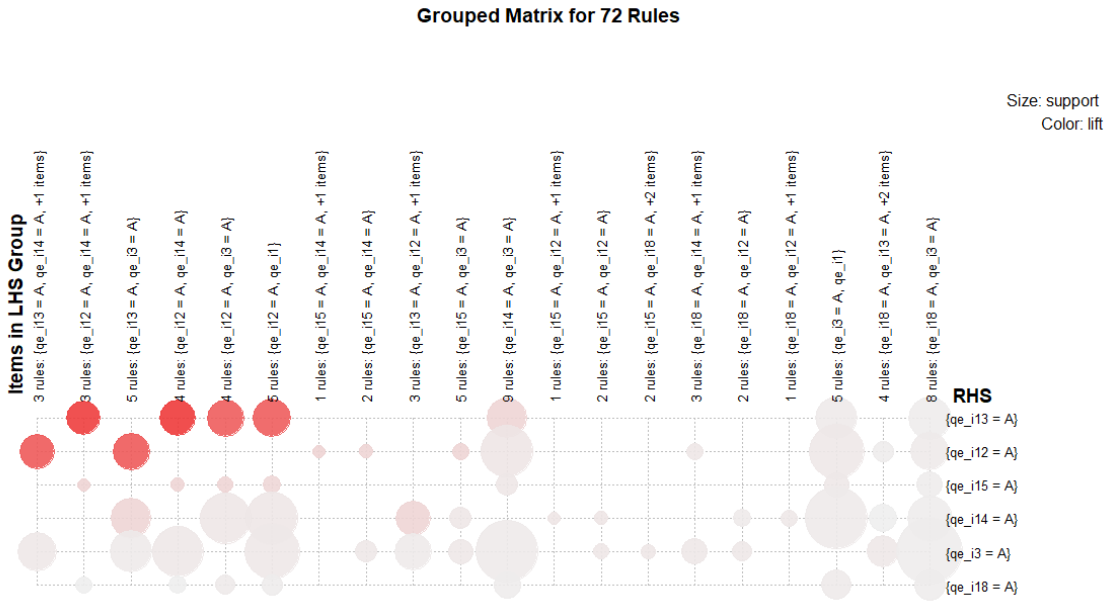

Figura 3. Matriz de visualização para 72 regras no ano de 2015.

Com o mesmo método aplicado aos anos anteriores, foi gerado o gráfico apresentado na figura 4, com as regras obtidas para o ano de 2016. Entre as regras geradas, uma infere que estudantes que afirmam ser brasileiros na questão 3 e concordam totalmente que o curso contribuiu para o desenvolvimento da sua consciência ética para o exercício profissional na questão 31, não participaram de programas de ensino no exterior, conforme suas respostas na questão 14, totalizando 78% dos estudantes que responderam o questionário.

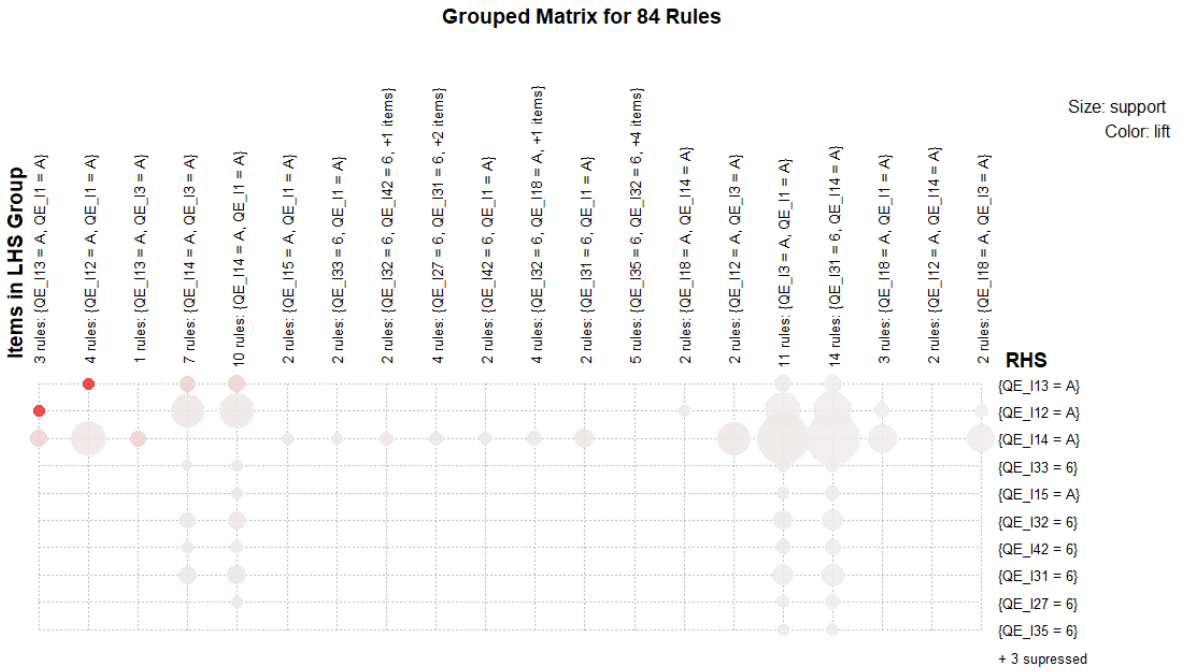

Figura 4. Matriz de visualização para 84 regras no ano de 2016.

As regras geradas pelo algoritmo utilizadas nos exemplos das apresentações dos anos de 2014, 2015 e 2016 são apresentadas na tabela 1 a seguir.

Tabela 1. Regras de exemplo para os anos analisados.

-	LHS	RHS	Support	Confidence
Ex. 2014	{qe_i14 = a,qe_i17 = a}	{qe_i3 = a}	0.7964280	0.9890552
Ex. 2015	{qe_i12 = a, qe_i14 = a, qe_i3 = a}	{qe_i18 = a}	0.7777144	0.8406557
Ex. 2016	{qe_i3 = A, qe_i31 = 6}	{qe_i14 = a}	0.7871561	0.98049394

5. Conclusões

Com a execução das etapas da metodologia CRISP-DM com os dados fornecidos pelo Inep para a realização deste projeto de mineração, foi possível a visualização de um grande conjunto de regras a respeito do Questionário do Estudante respondido pelos discentes das IES do Norte do Brasil ao realizar o ENADE.

Houve necessidade da limpeza, transformação e adequação dos dados para adequação ao algoritmo de Apriori, escolhido para realização da tarefa de associação dos atributos.

As regras geradas, juntamente com os gráficos executados no software R foram de fácil entendimento e de grande valia para a interpretação dos resultados, o que possibilitou o melhor entendimento dos resultados.

Como destaque da mineração tem-se a identificação de que os estudantes de IES da região Norte do Brasil carecem de maiores oportunidades com relação ao desenvolvimento atividades em outros países e assim obter experiências que contribuam para sua formação profissional. Também foi verificado que mais da metade desses estudantes tem queixas relativas à falta de apoio na sua formação acadêmica.

Como conclusão desse projeto de mineração tem-se que o processo foi de grande importância para a análise da massa de dados existente e até então pouco explorada. As regras geradas pela aplicação do algoritmo de Apriori foram iniciais e precisam de análises mais profundas. Como trabalho futuro, pretende-se analisar as demais regras geradas, buscando com isso identificar mais oportunidades de melhorias que possam ser implementadas pelas IES analisadas.

Vale ressaltar que os dados utilizados neste trabalho para sua realização são públicos e de livre acesso, fazendo com que os padrões definidos nesse trabalho possam ser replicados por outros grupos de Instituições de Ensino Superior do Brasil.

Referências

AGRAWAL, Rakesh; IMIELINSKI, T.; SWAMI, A. **Mining Association Rules between Sets of Items in Large Databases.**SIGMOD ,Washington,USA, 1993. Disponível em: <https://dl.acm.org/citation.cfm?id=170072>. Acesso em: 22 abr. 2018.

Camilo, C., Silva, J. **Mineração de Dados: Conceitos, tarefas, métodos e ferramentas**. [S.l: s.n.], 2009. Disponível em: <http://www.portal.inf.ufg.br/sites/default/files/uploads/relatorios-tecnicos/RT-INF_001-09.pdf>. Acesso: 15 mar. 2018.

Chapman, P., et al. 2005**. CRISP-DM 1.0**. *The Modeling Agency.* 2005. Disponível em: <https://www.the-modeling-agency.com/crisp-dm.pdf.>. Acesso em: 6 mar. 2018.

Côrtes, S. da C., Porcaro, R. M.; Lifschitz, S. **Mineração de Dados – Funcionalidades, Técnicas e Abordagens**. PUC, 2002. Disponível em: <ftp://obaluae.inf.puc-rio.br/pub/docs/techreports/02_10_cortes.pdf>; Acesso em: 18 abr. 2018.

ISSN: 2447-0767

Da Silva, L. A., Peres, S. M. e Boscarioli, C. 2016.***Introdução à Mineração de dados com aplicações em R***. Rio de Janeiro : Elsevier, 2016. p. 22. 978-85-352-8447-8.

Hoed, R. M. **Análise da evasão em cursos superiores: o caso da evasão em cursos superiores da área de Computação.**Dissetação PPCA, 2016. Disponível em: <repositorio.unb.br/handle/10482/22575>. Acesso em: 12 mar 2018.

IBM. **Visão geral da ajuda do CRISP-DM**. *IBM Knowledge Center.* IBM, 2017. Disponível em: <https://www.ibm.com/support/knowledgecenter/pt-br/SS3RA7_17.1.0/modeler_crispdm_ddita/clementine/crisp_help/crisp_overview.html.> Acesso em: 10 mar 2018.

Jang, S. P. Park, K. H. Kim, Y. L. Cho, H. N. Yoon, T. S. **Comparison of H5N1, H5N8, and H3N2 Using Decision Tree and Apriori Algorithm**. Journal of Biosciences and Medicines, 2015, 3, 49-53. Disponívelem: <www.scirp.org/journal/PaperInformation.aspx?paperID=58431>Acessoem: 23 abr.2018.

Kalgotra, P., Sharda, R. **Progression analysis of signals: Extending CRISP-DM to stream analytics.** Big Data (Big Data), 2016 IEEE International Conference on. Disponívelem: <http://ieeexplore.ieee.org/document/7840937/>. Acessoem: 12 abr. 2018.

Kim, C. woo., Ahn, Se H., Yoon, T. **Comparison of Flavivirus Using Datamining-Apriori, K-means, and Decision Tree Algorithm. Comparison of Flavivirus Using Datamining-Apriori, K-means, and Decision Tree Algorithm.**Disponível em: <ieeexplore.ieee.org/iel7/7885467/7890033/07890130.pdf> Acesso em: 22 abr. 2018.

Kumar, V., Chadha, A. **Mining Association Rules in Student's Assessment Data.** IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 5, No 3, September 2012. Disponível em: <http://citeseerx.ist.psu.edu/> . Acesso em: 30 mar 2018.

Leão H.A.T., Canedo E.D., Ladeira M., Fagundes F. (2018) **Mining ENADE Data from the Ulbra Network Institution.** In: Latifi S. (eds) Information Technology - New Generations. Advances in Intelligent Systems and Computing, vol 738. Springer, Cham. Disponível em: <https://link.springer.com/chapter/10.1007%2F978-3-319-77028-4_39#citeas>. Acesso em: 23 abr. 2018.

Liberotto, S. R., Mozzaquatro, P. M. **Análise dos algoritmos de mineração J48 e apriori aplicados na detecção de indicadores de qualidade de vida e saúde**. Revista Interdisciplinar de Ensino, Pesquisa e Extensão, vol 1, 2013. Disponível em: <revistaeletronica.unicruz.edu.br/index.php/eletronica/article/view/26-37>. Acessoem: 19 abr. 2018.

Luna, J. M., Padillo, F., Pechenizkiy, M. and Ventura, S. **Apriori Versions Based on MapReduce for Mining Frequent Patterns on Big Data.** IEEE TRANSACTIONS ON CYBERNETICS, 2017. Disponívelem: <ieeexplore.ieee.org/document/8052219/>Acessoem: 22 abr. 2018.

Mobasher, G. Shawish, A. Ibrahim, O. **Educational Data Mining Rule based Recommender Systems**. Proceedings of the 9th International Conference on Computer Supported Education, 292-299, 2017, Porto, Portugal. Disponível em: <http://www.scitepress.org/DigitalLibrary/>. Acesso em: 23 abr.2018.

Moro, S., Laureano, R. M. S., Cortez P. **USING DATA MINING FOR BANK DIRECT MARKETING: AN APPLICATION OF THE CRISP-DM METHODOLOGY**. 2011. Disponívelem<https://repositorium.sdum.uminho.pt/handle/1822/14838>. Acessoem: 23 de mar. 2018.

Olson, D. L.; Delen, D. **Advanced Data Mining Techniques**. Berlin: Springer, 2008. Disponível em: <https://pdfs.semanticscholar.org/c1c7/4829d6430d468a1fe1f75eae217325253baf.pdf>. Acesso em: 13 abr. 2018.

Ougiaroglou, S., Paschalis, G., **Association Rules Mining from the Educational Data of ESOG Web-Based Application.** AIAI 2012: Artificial Intelligence Applications and Innovations pp 105-114. Disponívelem: <https://link.springer.com/chapter/>. Acesso em: 27 mar 2018.

Tan, P., Steinbach, M., Kumar, V. **Introdução ao DATAMINING Mineração de Dados**. Ciência Mo ed. Rio de Janeiro: CiênciaModernaLtda, 2009.

ISSN: 2447-0767