O Uso de Ferramentas de Mineração de dados como auxílio na Prevenção da Evasão nas Universidades

Gabriel Rosa, Isaac Pimentel Fernandes Sobrinho, Diego Rodrigues

Curso de Licenciatura em Computação – Instituto Federal de Educação, Ciência e Tecnologia do Tocantins (IFTO) – Campus de Dianópolis, 77300-000 – Dianópolis – TO – Brasil

gabriel.lcc.rosa@gmail.com, isaacpimentelf@gmail.com, diego.rodrigues@ifto.edu.br

Resumo: Este documento consiste em apresentar brevemente os conceitos sobre mineração de dados e suas etapas de desenvolvimento, visando buscar métodos de auxílios para gestores e coordenadores, a fim de diminuir a evasão dos alunos em cursos de graduação nas instituições superiores. O desenvolvimento deste projeto fundamentou-se na pesquisa feita pelos autores deste artigo, que após a aplicação de um questionário no Instituto Federal de Educação, Ciência e Tecnologia do Tocantins, no curso de Licenciatura em Computação no Período noturno, obtiveram dados que orientam a utilização da ferramenta que auxilia a desenvolver padrões, regras de associação e arvore de classificação, através de algoritmos específicos.

1. Introdução

> As áreas governamentais, corporativas e científicas têm promovido um crescimento explosivo em seus bancos de dados, superando em muito a usual capacidade de interpretar e examinar estes dados, gerando a necessidade de novas ferramentas e técnicas para análise automática e inteligente de bancos de dados [FAYYAD et al. 1996].

Mineração de dados é um processo de tratamento de grandes quantidades de dados para encontrar padrões e correlações, ou seja, consiste de técnicas para extração de dados. Mineração de dados ou Data Mining é o termo que passou a ser adotado em 1990. Em sua composição contém disciplinas que caminham entrelaçadas: Estatística (coleta e análise de dados que obtém resultados numéricos), inteligência artificial e machine (algoritmos que permitem o computador aprender).

Weka (Waikato Environment for Knowledge Analysis) é um software livre desenvolvido por pesquisadores em 1993 na Nova Zelândia para uso de mineração de dados. A capacidade de realizar tarefas de Classificação é uma de suas principais vantagens, é também capaz de minerar regras de associação e clusters de dados e sua principal aplicabilidade é no meio acadêmico. Weka é uma ferramenta flexível e pode ser utilizada no modo de interface gráfica ou no modo console.

ISSN 2179-5649

Apriori foi proposto pela equipe de pesquisa da IBM - Projeto QUEST em 1994, o algoritmo serve para encontrar padrões associativos, ou seja, regras de associação. O Algoritmo J48 é utilizada para classificação dos dados, ela facilita a criação de modelos em arvore.

Essas ferramentas auxiliam a desenvolver pesquisas e competências de manipulação. Na prática do uso das ferramentas apresentadas, o estudo realizado pelos autores deste artigo tem como base a ideia de identificação dos problemas de evasão dos alunos nas Universidades Acadêmicas, analisando criticamente os pontos e dados coletados em turmas do curso de Licenciatura em Computação no Instituto Federal de Educação, Ciência e Tecnologia do Tocantins, campus Dianópolis, utilizando os softwares de mineração de dados citados no contexto e apresentando resultados eficientes na aplicação.

2. Mineração de Dados

Nas últimas décadas a mineração de dados cresceu exponencialmente, aumentando significativamente a quantidade de dados legíveis por máquinas na forma de arquivos e bancos de dados. Pode-se definir mineração de dados como descoberta de novas informações, em termos de padrão ou regras com base em grandes quantidades de dados. Deve ser aplicada de modo eficiente em grandes arquivos e bancos de dados para melhorar sua utilidade.

Em relatórios publicados pela Gartner Report, a mineração de dados tem sido aclamada como uma das principais tecnologias para o futuro próximo. O processo de **descoberta do conhecimento nos bancos de dados**, também conhecido como **KDD** (***Knowledge Discovery in Databases***), pode ser aplicado a qualquer problema de identificação de padrões em dados e contém fundamentação de diversas áreas, como: probabilidade, estatística, inteligência artificial, banco de dados e visualização de dados.

Em suma, dados podem ser definidos como um conjunto de fatos sobre determinado assunto e padrões como uma linguagem ou modelo que descreve um subconjunto destes fatos, ou seja, identificar um padrão é ajustar um modelo aos dados ou identificar uma estrutura no conjunto de dados, descrevendo-os de forma genérica.

Existem seis etapas a serem executadas antes da descoberta de um novo conhecimento:

- Seleção de dados;
- Limpeza de dados;
- Enriquecimento;
- Transformação ou codificação de dados;
- Mineração de dados;
- Relatório e exibição das informações.

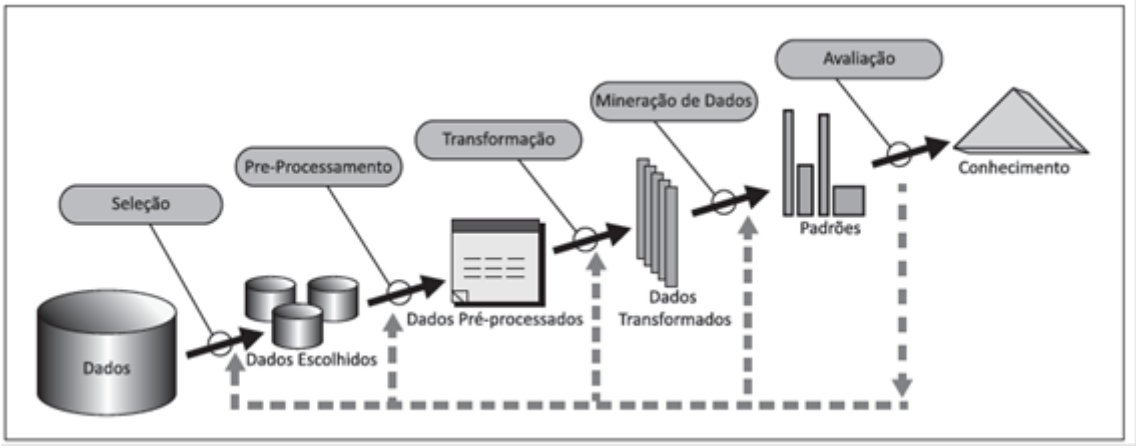

Figura 1. Etapas da descoberta do conhecimento. (FAYYAD et al. 1996)
Fonte. Educ. rev. vol.32 no.1 Belo Horizonte Jan./Mar. 2016.

Como exemplo, considere o banco de dados de transformação mantido pelos autores deste trabalho. Os dados coletados entre os alunos do curso superior de licenciatura em computação continham as seguintes informações: período de curso, faixa etária, relação de estudo fora da instituição de ensino, se exerce alguma atividade remunerada, se possui alguma formação técnica/superior, e se já reprovou em alguma matéria do curso.

Durante a Seleção de dados, dados sobre faixa etária ou formação técnica/superior podem ser relacionadas, ou estudo e aprovação pode ser selecionada entre outras inúmeras possibilidades de associação. O processo de Limpeza de dados, então, pode corrigir informações inválidas e eliminar registros incorretos. O Enriquecimento melhora os dados com fontes de informação adicional para anexar em cada registro. A Transformação ou Codificação de dados é feita para reduzir a quantidade de dados. Portanto, somente depois de pré-processados, as técnicas de Mineração de dados são usadas para descobrir diferentes regras e padrões.

> A fase de mineração de dados é uma fase do processo de descoberta de conhecimento em banco de dado. Esta etapa é responsável pela aplicação dos algoritmos que são capazes de identificar e extrair padrões relevantes presente em banco de dados (HAN,2001; WITTEN, 2000).

Através do resultado obtido, é possível descobrir novas informações, são elas: regra de associação, padrões sequenciais e arvore de classificação.

Tanto a mineração de dados, quanto a descoberta de conhecimento possuem alguns objetivos e aplicações finais. Esses objetivos, de modo geral, estão inseridos nas seguintes classes: previsão, identificação, classificação e otimização. O processo de Descoberta de Conhecimento em banco de dados é um processo não trivial de identificação de padrões novos, válidos e potencialmente úteis (FAYYAD et al.; 1996).

3. O que é Weka?

Weka é uma um software livre de Data Mining do tipo open source. Foi desenvolvido na Universidade de Waikato — Nova Zelândia, utiliza linguagem Java e foi desenvolvido dentro das especificidades da **GPL** (**General Public Licence**). É a ferramenta de mineração de dados mais utilizada no âmbito acadêmico e sua principal função é a tarefa de classificação, seus outros métodos também são bastante utilizados, especialmente a mineração de regras de associação.

A ferramenta tem uma interface gráfica simples e de fácil acesso, que permite a execução de algoritmos de forma interativa e colaborativa para com o usuário.

Os principais métodos aplicados pela ferramenta Weka:

- Métodos de classificação;
- Métodos para predição numérica;
- Métodos de agrupamento;
- Métodos de associação.

Sua formatação é a partir de arquivos com entradas no formato "ARFF" (Attribute Relation File Format), correspondente a um arquivo de texto que obtém um aglomerado de informações, precedido por um cabeçalho que é utilizado para importar as informações aos campos que compõem os conjuntos de informações.

4. Algoritmos

4.1 Apriori

As regras de associação têm uma inestimável importância para as tarefas de mineração de dados. O algoritmo Apriori é o mais utilizado no âmbito acadêmico para descobrir regras de associação. Foi apresentado inicialmente em **Mining association rules between sets of items in large databases** (AGRAWAL, et. Al, 1993), e consiste em descobrir padrões associativos. Possui capacidade de executar um grande número de atributos, extraindo como resultado inúmeras alternativas combinatórias. Seu desempenho é excelente e adota como objetivo encontrar regras de associação para todas as expressões possíveis e entender tendências que possam ser usadas para explorar padrões de comportamento e de dados.

4.2 J48

O classificador por arvore de decisão é uma técnica bastante utilizada para a classificação, pois é a indução descendente de arvores de decisão. O algoritmo J48 está presente na ferramenta de mineração de dados Weka, e é muito utilizado para produzir arvore de fácil entendimento, que as pessoas possam compreender de uma forma simplificada.

> Neste sentido, a mineração de dados aparece como uma metodologia auxiliar, que otimiza o processo de escolha de descritores e limiares das classes, pois procura definir padrões, associações, mudanças, anomalias e estruturas significativas entre os dados. Essa técnica extrai informações de uma determinada base de dados, criada por meio da tarefa de

classificação e da técnica de árvores de decisão. Ficando a cargo do analista apenas os processos de elaboração da rede hierárquica, segmentação e coleta de amostras (SOUZA, 2012).

5. Objetivos

5.1 Objetivo Geral

Proporcionar conhecimentos válidos para que gestores de instituições consigam identificar o padrão dos alunos através de índices de reprovação, e associar a medidas de intervenção para diminuir a evasão dos alunos nos cursos superiores.

5.2 Objetivos Específicos

- Definir padrões de alunos através do seu rendimento;
- Gerar novas informações para auxiliar a coordenação dos cursos superiores;
- Desenvolver medidas a fim de reduzir a evasão das instituições de nível superior.

6. Metodologia

Demonstraremos neste relato de experiência, o percurso de desenvolvimento da pesquisa exploratória, de abordagem quantitativa e qualitativa, em que utilizamos como instrumento de geração de dados a aplicação do questionário proposto pelo professor Diego Rodrigues e desenvolvido pelos autores desta obra. O questionário tem como propósito identificar padrões nos alunos do Instituto Federal de Educação, Ciência e Tecnologia do Tocantins, campus Dianópolis. O roteiro das questões aplicadas na pesquisa foi composto pelas seguintes perguntas:

Tabela 1. Modelo do questionário preenchido pelos alunos.

1. Qual período você cursa?	2º( ) 3°( ) 4º( ) 5°( )
2. Qual ano do seu nascimento?	
3. Você trabalha?	Sim( ) Não( )
4. Você estuda fora da universidade?	Sim( ) Não( )
5. Você já reprovou em alguma matéria?	Sim( ) Não( )
6. Você tem alguma formação técnica/superior?	Sim( ) Não( )

De acordo com o Estatuto da Juventude, a faixa etária dos jovens é de 15 a 29 anos de idade, e com base em um tratado da ONU, a faixa etária dos idosos nos países subdesenvolvidos é a partir dos 60 anos, consideramos então a faixa etária dos adultos entre 30 e 59 anos de idade.

Para a aplicação do questionário contamos com os alunos do curso superior de Licenciatura em Computação entre a faixa etária jovem, adulto e idoso, no período

noturno. Os dados foram obtidos por meio do preenchimento de 63 questionários impressos, respondidos pelos alunos que concordaram em participar da pesquisa.

Registradas conforme as questões apresentadas, as respostas foram organizadas em planilha do *Excel* para a tabulação e análise dos dados, separando-se por períodos: 2°, 3º, 4º e 5º. Após a tabulação, as questões foram transcritas para um documento *WordPad* com cabeçalho indicando as perguntas do questionário e as respostas sendo introduzidas na sequência, a fim de salvar o documento no formato de arquivo "ARFF" e ser interpretado pela ferramenta de mineração de dados Weka. Ao inicializar a ferramenta Weka e incluir o arquivo salvo no formato "ARFF", obtivemos os resultados que serão discutidos a seguir.

7. Resultados

Para realizar as análises de dados coletados, foram levados em consideração alguns aspectos sobressalentes. Optamos por agrupar as respostas por temas relacionáveis. Na nossa amostra, portanto, foram analisados os 63 questionários preenchidos que geraram inúmeras regras de associação.

Dessa forma, os dados nos quais nos concentramos para realizar as discussões deste artigo expressam quatro características relacionais que se destacam nas respostas dos alunos e evidenciam seus padrões por: Faixa etária, Trabalho, Estudo e Reprovação.

Tabela 2. Trabalho relacionado à Reprovação.

Antecedente	Consequente	Confiança
Reprovou = Sim; Formado = Sim	Trabalha = Sim	0.93
Trabalha = Sim; Formado = Não	Reprovou = Sim	0.82

Pode-se observar através da tabela acima que, existe uma evidente relação entre Trabalho e Reprovação ou Reprovação e trabalho, pois, as regras selecionadas se confirmam e possuem mais de 80% de confiança e uma média de 87,5%.

Tabela 3. Trabalho relacionado a Estudo.

Antecedente	Consequente	Confiança
Trabalha = Não; Reprovou = Não	Estuda = Sim	1
Faixa etária = Jovem; Estuda = Sim	Trabalha = Não	0.84
Estuda = Não	Trabalha = Sim	0.73

Já nesta tabela podemos notar que, pessoas que exercem algum tipo de atividade trabalhista não são habituadas a estudar, ou por falta de tempo ou condições não favoráveis em seu âmbito de trabalho. Como mostrado, a confiança reside acima dos 70% e a média é de 85,6%.

Tabela 4. Faixa etária relacionada à Reprovação.

Antecedente	Consequente	Confiança
Trabalha = Não; Reprovou = Sim	Faixa etária = Jovem	1
Trabalha = Não; Reprovou = Sim; Formado = Não	Faixa etária = Jovem	1
Faixa etária = Jovem; Estuda = Não	Reprovou = Sim	0.92

Apesar da faixa etária Jovem ser a maior ocupante de vagas do curso de Licenciatura em Computação no Instituto Federal de Educação, Ciências e Tecnologia do Tocantins - campus Dianópolis, é também o grupo com maior índice de reprovação. Como a tabela aponta, a confiança dos dados associados é acima dos 90%, e a média é de 97,3%.

Tabela 5. Estudo relacionado à Reprovação.

Antecedente	Consequente	Confiança
Reprovou = Não	Estuda = Sim	0.89
Estuda = Não	Reprovou = Sim	0.87

Nota-se através desta tabela que a relação de Estudo com Reprovação é proeminente e estão atreladas ao êxito. Pelo fato da confiança estar acima de 85% e possuir uma média de 88%.

8. Discussões

Ao analisar criticamente todo o conhecimento gerado pela pesquisa, adquirimos informações válidas para colocar em pauta. A discussão preludio se fez para examinar minunciosamente os resultados e alcançar o objetivo principal do artigo em auxiliar gestores e coordenadores de curso a identificar os padrões dos alunos, e tomar medidas preventivas para evitar a evasão das instituições de ensino superior.

A discussão ao atingir seu momento culminante, pôde identificar características de alunos que tendem a desistir dos cursos de graduação. Para identificar esses padrões foi levado em consideração aspectos econômicos, familiares e educacionais.

- Econômico: muitas vezes a falta de verba faz com que estudantes abandonem a universidade, devido aos custos de alimentação, transporte, moradia e outros fatores de necessidade básica.
- Familiar: nas últimas décadas o modelo familiar comum sofreu algumas alterações, mas de todo modo ainda existe famílias que suprimem os desejos dos filhos ou dos mais jovens. Isso geralmente ocorre por dois motivos, o primeiro, é quando a família obriga a pessoa a entrar em um curso superior para realizar o desejo dos pais que não tiveram oportunidade de estudar na época de juventude e não possuem interesse de ingressar na faculdade a essa altura da vida. O segundo motivo é quando a família pressiona o sujeito para seguir um legado como existe famílias com varais gerações de médicos, advogados entre outros seguimentos profissionais.
- Educacional: pelo fato de muitos possuírem um déficit na educação básica, é comum a desistência nas universidades por acharem que não vão conseguir adquirir o conhecimento. Ou então, por não adquirirem um conhecimento sólido nos períodos iniciais, ao decorrer do curso com as reprovações em algumas matérias gera um descontrole psicológico que leva a desistência do curso.

Já existem medidas para ajudar o estudante em sua jornada acadêmica. Mas esse trabalho deve ser intensificado e divulgado entre todos os alunos em situações de vulnerabilidade. Como exemplo podemos citar os auxílios financeiros, apoio psicológico, palestras de incentivos, atendimento com professores ou monitorias, apoio a formação de grupos de estudos, disponibilização de ambientes estruturados para estudos coletivos em turno oposto ao de aula, é importante também que os alunos dos períodos avançados disponibilizem tempo para ajudar os alunos dos períodos iniciais, que ao entrarem na universidade se depara com coisas que para eles, parecem ser de "outro mundo". Essas ações visam sanar a precariedade em que o estudante se insere, fazendo com que o aluno se sinta seguro, para que haja estímulo e motivação que venha a trazer o interesse de continuar na universidade e possa concluir o curso.

9. Considerações Finais

A proposta do artigo está alinhada em apresentar como a utilização das técnicas de mineração de dados em ambientes educacionais pode ser uma ferramenta para municiar a gestão de uma instituição de ensino nos quesitos de gerenciamento, tomada rápida de decisões para resgatar os alunos com dificuldades e intervir na vida acadêmica ao notar um baixo rendimento através de padrões associados as notas semestrais.

Esta pesquisa se propôs, com o objetivo geral voltado aos coordenadores e diretores de Universidades, visando elaborar propostas de intervenções para que o índice de desistência e reprovação seja diminuído das universidades superiores.

Através do método da pesquisa podemos extrair informações que norteiam esta constante evasão. Com o uso de ferramentas que auxiliam o estudo e a mineração de dados, obtivemos o embasamento para ideias e sugestões que poderão servir de auxílio.

Após a análise dos dados e as discussões dos resultados, chegamos à conclusão de que os alunos que exercem alguma atividade remunerada diariamente, não tem tempo suficiente para estudar fora do período da universidade como os alunos que não trabalham, isso faz com que os índices de reprovação se elevem nessa categoria. E é obviamente constatado que a reprovação é a consequência da falta de estudo, o conhecimento passado pelo docente em sala não é o suficiente para que o discente se torne profissional, mas sim a busca de informações em outras fontes e a prática dos estudos.

Outra hipótese de grande importância que pode explicar o fato de constante desistência são as características dos alunos que desistem por aspectos econômicos, familiares e educacionais; ou seja, é indispensável o estudo diário em casa, porém existem também outras barreiras que nem todos os alunos estão encorajados a enfrentar.

Através das análises, foi possível extrair algumas recomendações de práticas a serem seguidas pelos gestores, são as medidas que devem usar para controlar a evasão, como auxílios financeiros, estímulo ao estudo, apoio aos alunos a formarem grupos de estudos, períodos estudando de forma conjunta, ambientes estruturados para estudos fora do horário de aula, e outras medidas que servem de auxílio ao aluno. É importante que haja apoio aos alunos, e em ligação com o realismo pedagógico, cada um tem sua forma de aprender. Contudo, ficam essas recomendações aos diretores e coordenadores de cursos superiores, e como dizia Aristóteles – “A educação tem raízes amargas, mas os seus frutos são doces.”.

10. Referencias

SILVA, Marcelino Pereira dos Santos. Mineração de Dados - Conceitos, Aplicações e Experimentos com Weka. Rio Grande do Norte, Brasil, 2004. Disponível: http://www.lbd.dcc.ufmg.br/colecoes/erirjes/2004/004.pdf

Conceitos de mineração de dados. Disponível: https://msdn.microsoft.com/pt-br/library/ms174949.aspx

Mineração de dados no MySQL com a ferramenta Weka. Disponível: http://www.devmedia.com.br/mineracao-de-dados-no-mysql-com-a-ferramenta-weka/26360

Figura 1. Etapas da descoberta do conhecimento. (FAYYAD et al. 1996). Disponível: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0102-46982016000100133

FONSECA, Stella Oggioni. NAMEN, Anderson Amendoeira. MINERAÇÃO EM BASES DE DADOS DO INEP: UMA ANÁLISE EXPLORATÓRIA PARA NORTEAR MELHORIAS NO SISTEMA EDUCACIONAL BRASILEIRO. Rio de Janeiro, Brasil, 2016. Disponível: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0102-46982016000100133

Mineração de Regras de Associação com a Ferramenta de Data Mining Weka. Disponível: http://www.devmedia.com.br/mineracao-de-regras-de-associacao-com-a-ferramenta-de-data-mining-weka/20478

LIBRELOTTO, Solange Rubert. MOZZAQUATRO, Patricia Mariotto. ANÁLISE DOS ALGORITMOS DE MINERAÇÃO J48 E APRIORI APLICADOS NA DETECÇÃO DE INDICADORES DA QUALIDADE DE VIDA E SAÚDE. Revista Interdisciplinar Ensino, Pesquisa e Extensão. Vol.1, Nº1, 2013. Disponível: http://revistaeletronica.unicruz.edu.br/index.php/eletronica/article/view/26-37

CASAS, Pedro Henrique Bragioni Las. MINERAÇÃO DE DADOS APLICADA. Disponível: http://homepages.dcc.ufmg.br/~pedro.lascasas/aula_2_minera%C3%A7%C3%A3o_de_dados_aplicada_weka.pdf

WITTEN. Ian H. FRANK, Eibe. Data Mining Practical Machine Learning Tools and Techniques. Second Edicion, Morgan Kaufmann Publishers, 2005. Disponível: ftp://ftp.ingv.it/pub/manuela.sbarra/Data%20Mining%20Practical%20Machine%20Learning%20Tools%20and%20Techniques%20-%20WEKA.pdf

DAMASCENO, Marcelo. INTRODUÇÃO A MINERAÇÃO DE DADOS UTILIZANDO O WEKA. Rio Grande do Norte, Brasil, 2010. Disponível: http://connepi.ifal.edu.br/ocs/index.php/connepi/CONNEPI2010/paper/viewFile/258/207

NASCIMENTO, Lídice Cabral. CRUZ, Carla Bernadete Madureira. Rio de Janeiro, Brasil, 2013. Disponível: http://www.dsr.inpe.br/sbsr2013/files/p0789.pdf

SILBERSCATZ, Abraham. KORTH, Henry F. Sudarshan, S. Sistema de Banco de Dados. 5ª Edição, Elsevier, 2006.

ELMASRI. NAVATE. Sistema de Bancos de Dados. 6ª Edição, Pearson, 2011.