Previsão de Estudantes com Risco de Evasão Utilizando Técnicas de Mineração de Dados

Laci Mary Barbosa Manhães[1], Sérgio Manuel Serra da Cruz[2], Raimundo J. Macário Costa[1], Jorge Zavaleta[1], Geraldo Zimbrão[1]

[1] PESC/COPPE– Universidade Federal do Rio de Janeiro (UFRJ)
Caixa Postal 68.511 – 21941-972 – Rio de Janeiro – RJ – Brasil

[2] Instituto de Ciências Exatas – Universidade Federal Rural do Rio de Janeiro (UFRRJ)
Seropédica, RJ - Brasil

{manhaes, macario, jorgejzg, zimbrao}@cos.ufrj.br, serra@ufrrj.br

Abstract. *The universities are facing the problem of reducing the students drop out rate in undergraduate courses. The problem occurs in several Brazilians universities and generates prejudices to the country, students and universities. In this sense, our goal is to identify, in the early stage, students who may be at risk of dropping out in undergraduate courses. The data miming technique has many algorithms that can be used in order to identify student at risk to fail to complete the course. This research is based on academic data from undergraduate students of the Brazilian University - UFRJ. The results reveal that is possible to identify the final course situation of the freshman with 80% of precision using the first academic semester grades.*

Resumo. *As universidades enfrentam o desafio de reduzir os índices de evasão dos alunos nos cursos de graduação. O problema ocorre em diversas universidades brasileiras gerando prejuízos para o país, alunos e universidades. Neste sentido, nosso objetivo é identificar precocemente alunos em risco de evasão nos cursos de graduação. A técnica de mineração de dados oferece diversos algoritmos que podem ser empregados para identificar alunos em risco de abandono. Este trabalho apresenta um estudo utilizando dados acadêmicos de alunos de graduação de uma universidade brasileira-UFRJ. Os resultados mostraram que utilizando as primeiras notas semestrais dos calouros é possível identificar com precisão de 80% a situação final do aluno no curso.*

1. Introdução

A evasão e retenção nos cursos de graduação das universidades públicas brasileiras representam um problema complexo e atinge inúmeras instituições. Existe uma preocupação dos governos e das instituições em atenuar os índices de evasão dos cursos por parte dos estudantes universitários. O Governo Federal, através do Programa de Apoio a Planos de Reestruturação e Expansão das Universidades Federais – REUNI [Governo Federal 2007] relata que "os índices de evasão de estudantes nos cursos de graduação atingem, em alguns casos, níveis alarmantes". A meta do governo é a

elevação gradual da taxa de conclusão média dos cursos de graduação presenciais para noventa por cento.

Atualmente, as instituições de ensino superior oferecem a cada ano um crescente número de vagas para novos alunos ingressarem nos cursos de graduação. No entanto, partes dos alunos que entram na universidade não concluem o curso. A conquista de uma vaga em uma universidade pública seguido do abandono tornou-se um problema generalizado, independente da instituição [INEP 2009]. Os índices de evasão variam entre as universidades e entre os cursos. O estudo das causas da evasão e a tomada de medidas preventivas estão fortemente ligados ao contexto de cada instituição de ensino. Por exemplo, a identificação dos fatores que influenciam a evasão e atribuição de uma ordem de importância para estes fatores é um trabalho complexo que está diretamente ligado a análise do conjunto de alunos.

Os índices de evasão podem ser medidos em diversos níveis de abrangência, por exemplo, o índice de evasão em uma disciplina em um período específico [Hämäläinen et al. 2004], a evasão em um curso de graduação [Barroso e Falcão 2004; Soares 2006]. Existem várias condições que propiciam evasão. Por exemplo, no contexto institucional deve ser considerado que cada universidade possui características que atraem maior ou menor número de alunos em função de sua localização geográfica, público alvo, adequação dos cursos ao contexto sócio-econômico da região. No contexto do curso de graduação algumas condições são a atualidade do currículo do curso, sua adequação à formação para o mercado de trabalho e tempo de duração.

Segundo Barroso e Falcão (2004) as condições que motivam a evasão escolar são classificadas sob três agrupamentos: i) *econômica* - impossibilidade de permanecer no curso por questões sócio-econômicas; ii) *vocacional* – o aluno não se identifica com o curso; iii) *institucional* – abandono por fracasso nas disciplinas iniciais, deficiências prévias de conteúdos anteriores, inadequação aos métodos de estudo, dificuldades de relacionamento com colegas ou com membros da instituição.

Este trabalho tem como objetivo identificar precocemente o subconjunto dos alunos do curso de Engenharia da Escola Politécnica da UFRJ que apresentam risco de evasão utilizando um conjunto de técnicas de mineração de dados. Atualmente, o processo de identificação é manual, subjetivo, empírico e sujeito a falhas, pois depende primordialmente da experiência acadêmica e do envolvimento dos docentes. Geralmente, estes desempenham inúmeras atividades além das atividades de sala de aula, portanto é difícil acompanhar e reconhecer as necessidades de cada aluno e identificar aqueles alunos que apresentam risco de evasão. Portanto, a adoção de mecanismos automatizados que viabilizam não só a detecção precoce de grupos de alunos com risco de evasão é uma condição importante para reduzir o problema da evasão.

Este trabalho está estruturado da seguinte forma: Na seção 2 discutem-se os trabalhos relacionados referentes ao problema da evasão escolar utilizando mineração de dados. Na seção 3, contextualizam-se o problema e a descrição da base de dados. Na seção 4 descrevem-se a ferramenta e os algoritmos de mineração de dados. Na seção 5 descrevem-se a aplicação dos algoritmos sobre os dados e relatam-se os três experimentos. No final do artigo, seções 6 apresentam-se uma conclusão do estudo e discussão sobre trabalhos futuros.

2. Trabalhos Relacionados

A previsão do desempenho acadêmico é um tema já estudado por diversos pesquisadores. Estudos mais antigos utilizam métodos estatísticos ou outros métodos para compreender o problema [Moore 1995; Johnston 1997; Davies 1997].

A utilização de técnicas de mineração de dados sobre dados educacionais é relativamente recente conforme destaca Baker et al. (2009, 2011) e Dekker et al. (2009). No entanto, a maioria dos trabalhos correlatos está restrita a identificar resultados em pequenos contextos relativos a apenas uma disciplina de curso não presencial.

Minaei-Bidgoli et al. (2006) utilizaram regras de associação para extrair padrões de informações em bases de dados geradas a partir de sistemas educacionais online. Os autores demostraram que um conjunto de regras permite identificar quais os atributos que caracterizam padrão de desempenho dos grupos de estudantes, neste caso, a pesquisa teve como base a disciplina de Física oferecida em ambiente on-line.

Hämäläinen et al. (2004) analisaram duas disciplinas de programação de computadores em um curso on-line. O trabalho utilizou regras de associação e modelos probabilísticos para identificar os fatores mais importantes para predizer os resultados finais nas duas disciplinas. Kotsiantis et al. (2003) compararam diversos algoritmos para detectar o mais adequado para predizer a evasão dos alunos.

Um trabalho mais abrangente foi realizado por Dekker et al. (2009), os autores analisaram dados dos alunos de graduação do curso presencial de Engenharia Elétrica da universidade de Eindhoven. Neste trabalho, identificou-se já no primeiro ano letivo os alunos com risco de evasão. Os autores avaliaram diversos algoritmos da ferramenta de mineração de dados Weka [Hall et al. 2009] afim de detectar o mais adequado. O experimento analisou diversos dados dos alunos e obteve entre 75 a 80% de acurácia com o classificador árvore de decisão.

A adoção de algoritmos de mineração de dados em dados educacionais para a previsão da situação acadêmica é um campo de investigação ainda não consolidado, necessita de investigações complementares tanto na definição dos atributos a serem utilizados quanto nas técnicas de mineração de dados empregadas [Castro et al. 2007, Baker 2009; Dekker et al. 2009]. Os autores, em linhas gerais, indicam pontos que precisam ser pesquisados para aprimorar a utilização da mineração de dados na identificação de estudantes com risco de evasão. Os principais pontos são: i) transformação dos dados (os dados colhidos nem sempre são diretamente tratados pelos algoritmos de mineração); ii) identificar os atributos mais relevantes; iii) identificar os algoritmos mais adequados; iv) aplicar os algoritmos para identificar outros grupos de estudantes.

3. Contextualização do problema e descrição da base de dados

A Escola Politécnica da Universidade Federal do Rio de Janeiro – UFRJ oferece cursos de graduação para formação de engenheiros em diversas áreas. Apesar da excelência na formação e da grande concorrência por uma vaga, o problema da evasão preocupa a direção da Escola [Saraiva e Masson 2003].

A base de dados utilizada neste trabalho foi diretamente coletada do sistema acadêmico da UFRJ, ela contém informações sobre os alunos de graduação que ingressaram no curso de Engenharia Civil da Escola Politécnica e suas cinco ênfases no

período de 1994 a 2005. Ressaltam-se os alunos tiveram a sua identificação preservada. A base de dados é composta por informações de 543 alunos que concluíram o curso de Engenharia e mais 344 registros de alunos que não concluíram. A base foi fracionada em duas classes distintas e bem definidas. A primeira classe composta por alunos que completaram todos os requisitos para aprovação e conclusão do curso. A segunda classe composta por alunos que não concluíram o curso por iniciativa própria (abandono ou trancamento de matrícula); ou por imposição da universidade (reprovação por nota, ultrapassar o prazo para conclusão do curso e sansão disciplinar).

Neste trabalho, o procedimento de seleção dos atributos ainda está em fase de estudos devido às limitações de acesso aos dados e pela falta de referencias sobre quais atributos são mais adequados para analisar o problema da evasão escolar utilizando mineração de dados. Sendo assim, baseado na informação que o maior número de evasões ocorre no início do curso, optou-se por utilizar dados acadêmicos dos períodos letivos que antecedem o maior número de evasões. Portanto, foram selecionadas as informações acadêmicas referentes ao primeiro período letivo [Saraiva e Masson 2003].

Identificaram-se as disciplinas mais cursadas relativas ao primeiro semestre, a saber: Introdução a Engenharia Civil, Engenharia e Meio Ambiente, Programação de Computadores I, Cálculo Diferencial e Integral I e Química. Uma vez identificada às disciplinas, foram selecionados 887 calouros que freqüentaram estas disciplinas. A base, então, foi composta pelas notas e a situação final (aprovado, reprovado por nota, reprovado por falta) em cada disciplina; o valor do coeficiente de rendimento acumulado no período e, por fim, o atributo identificador da classe do aluno.

O ganho de informação é utilizado para avaliar o quanto um atributo influencia o critério de classificação do algoritmo [Han 2005]. Seguindo a ordem do mais importante para o menos importante, os atributos da base foram ordenados como segue: coeficiente de rendimento do período, nota da disciplina de Cálculo Diferencial e Integral I, situação (Aprovado, Reprovado por Nota ou Reprovado por falta) na disciplina de Cálculo Diferencial e Integral I, notas nas disciplinas: Engenharia de Meio Ambiente, Programação de Computadores, Química, Introdução a Engenharia Civil, situação das disciplinas: Introdução a Engenharia Civil, Química, Programação de Computadores, Engenharia de Meio Ambiente.

4. Ferramenta e algoritmos de mineração de dados

Atualmente a aplicação das técnicas de mineração de dados é facilitada devido a existência de ferramentas que dispõem de recursos de análise de dados e implementam algoritmos específicos. A ferramenta de mineração de dados Weka [Hall et al. 2009] foi a escolhida para este trabalho devido a fatores: facilidade de aquisição e disponibilidade para *download* diretamente da página do desenvolvedor sem custo de utilização; presença de várias versões de algoritmos empregados na mineração de dados e disponibilidade recursos estatísticos para comparar o desempenho dos algoritmos.

4.1. Ambientes de Experimentos da ferramenta Weka

A ferramenta Weka disponibiliza o ambiente *Weka Experiment Environment* (WEE), este ambiente é apropriado para realizar comparações entre o desempenho de vários algoritmos de mineração de dados [Bouckaert et al. 2010]. O WEE permite selecionar

um ou mais algoritmos disponíveis na ferramenta e analisar os resultados de modo a identificar se um classificador é, estatisticamente, melhor do que os demais.

O WEE oferece três opções de estratificação da base de dados: i) *Cross-validation (default),* ii) *Train/Test Percentage Split (data randomized)* e iii) *Train/Test Percentage Split (order preserved)*. Para obter significância estatística nos desempenhos dos algoritmos, o ambiente foi configurado com um número padrão de execuções. Por exemplo, cada algoritmo é executado 10 vezes e seu desempenho final é obtido a partir da média das execuções. No caso do *10-fold cross-validation* significa que um classificador foi executado 100 vezes para os conjuntos de treinamento e teste.

O termo significância estatística refere-se à diferença entre os resultados de cada um dos algoritmos em relação a um algoritmo escolhido como base (*baseline*). O padrão de configuração do teste utiliza comparação *pair-wise T-Test* com significância de 5%.

O outro ambiente disponível na ferramenta Weka é o *Weka Explorer* (WE), este ambiente permite a seleção e execução de um algoritmo classificador por vez [Bouckaert et al. 2010]. Sendo assim, o resultado do experimento representa a acurácia de uma rodada do algoritmo. A comparação dos resultados dos classificadores não é efetuada de forma automática como no ambiente WEE.

O WE oferece quatro opções de estratificação da base de dados: *use training set, supplied test set, cross-validation* e *percentage split*. A opção *supplied test set* permite, diretamente, especificar o conjunto de teste, ou seja, fornecer para a ferramenta um conjunto sem rótulo de classe. A importância desta opção está na facilidade de selecionar os exemplos que vão compor o conjunto de treinamento e testes.

4.2. Algoritmos de mineração de dados

Selecionam-se 10 algoritmos de classificação disponíveis na ferramenta Weka, a escolha deve-se ao fato da ampla utilização dos algoritmos em diversos contextos [Wu et al. 2008; Witten et al. 2011; Dekker et al. 2009]. Os algoritmos selecionados são OneR (OR), JRip (JR), DecisionTable (DT), SimpleCart (SC), J48 (J48), RandomForest (RF), SimpleLogistic (SL), MultilayerPerceptron (MP), NaiveBayes (NB), BayesNet (BN). Apresentam-se os métodos de classificação empregados pelos algoritmos: aprendizado de regras (OneR e JRip), tabela de decisão (DecisionTable), árvore de decisão (SimpleCart, J48 e RandomForest), modelos lineares de regressão logística (SimpleLogistic), modelo de rede neural artificial (MultilayerPerceptron), modelos probabilístico (BayesNet), classificador probabilístico simples baseado na aplicação do teorema de Bayes (NaiveBayes).

5. Experimentos com algoritmos de mineração de dados

Neste estudo, realizaram-se três experimentos com o objetivo de comparar o desempenho dos algoritmos de mineração de dados aplicados ao domínio do problema. A análise dos resultados serve para identificar quais algoritmos são mais adequados para mineração de dados educacionais.

O algoritmo OneR foi escolhido como base de referência (baseline*),* a escolha do algoritmo *baseline* é experimental, optou-se pelo OneR por ser um classificador

muito simples, por utilizar um método de classificação de custo reduzido e obter uma acurácia alta [Han e Kamber 2006; Carvalho 2005; Witten et al. 2011].

5.1. Experimento 1

O primeiro experimento foi executado no ambiente WEE, foram selecionados dez algoritmos classificadores (listados na seção 4.2). A base de dados foi dividida em 10 conjuntos utilizando o método de validação cruzada (10 folds cross-validation). Os algoritmos, aplicados a base de dados, foram executados 10 vezes, valor padrão de configuração do ambiente.

Resultados do experimento 1: A ferramenta calculou a média das acurácias obtidas em cada rodada dos classificadores, Tabela 1.

Tabela 1. Acurácia média dos classificadores utilizando validação cruzada

Classificadores	OR	JR	DT	SC	J48	RF	SL	MP	NB	BN
Acurácia	78,39	77,88	78,07	78,92	77,86	76,74	78,32	76,36	78,85	78,78

A ferramenta mostrou que nenhum dos algoritmos utilizados neste experimento foi significativa melhor ou pior do que o OneR. De fato, todos os algoritmos testados apresentaram valores médios muito próximos, entre 76,36 a 78,92.

5.2. Experimento 2

Este experimento também foi realizado no ambiente WEE, utilizou-se a base de dados e os algoritmos da seção 4.2. A configuração do ambiente foi modificada para contemplar outra forma de dividir a base de dados, *Train/Test Percentage Split (data randomized),* esta opção utiliza um processo randômico para selecionar os exemplos dos conjuntos de treinamento e teste. O padrão da ferramenta é utilizar 66% dos registros para o conjunto de treinamento e 34% para teste [Bouckaert et al. 2010].

O objetivo deste experimento é verificar se a alteração na forma de seleção do conjunto de treinamento e teste afeta o desempenho dos classificadores.

Resultados Experimento 2: A ferramenta calculou a média das acurácias obtidas nas 10 execuções dos classificadores (Tabela 2). A ferramenta mostrou que nenhum dos algoritmos utilizados neste experimento foi significativa diferente do *baseline*.

Tabela 2. Acurácia média dos classificadores com seleção randômica do conjunto de treinamento e teste

Classificadores	OR	JR	DT	SC	J48	RF	SL	MP	NB	BN
Acurácia	78,50	78,44	79,63	79,36	77,87	77,28	79,76	76,35	80,12	79,66

Os resultados das avaliações dos desempenhos mostraram média de acertos dos classificadores entre 76 a 80%. Isto é um forte indicativo que os atributos utilizados são suficientes para executar a previsão dos alunos com risco de evasão logo ao final do primeiro semestre letivo.

Comparando os resultados dos experimentos 1 e 2, verificam-se que as duas formas de divisão da base de dados e composição dos conjuntos de treinamento e teste não são significativamente diferentes. Comparando os valores de cada classificador

disponível nas Tabelas 1 e 2, observam-se resultados muito próximos para maioria dos classificadores, a maior diferença de desempenho está abaixo de 2% para os algoritmos DecisionTable (DT), SimpleLogistic (SL) e NaiveBayes (NB). Os experimentos 1 e 2 foram refeitos modificando o número de execuções dos algoritmos: de 10 para 100 execuções sobre a base de dados, os novos valores obtidos foram similares aos anteriores.

5.3. Experimento 3

O terceiro experimento foi realizado no ambiente WE, os 10 algoritmos definidos na seção 4.2 foram individualmente carregados e executados. A Tabela 3 mostra a acurácia de uma execução dos algoritmos aplicados a base de dados. A opção *Supplied test set* foi escolhida como forma de seleção do conjunto de treinamento e teste. Esta opção permite que o usuário determine os conjuntos de treinamento e testes.

Neste experimento, o conjunto de treinamento é composto por 2/3 da base (599 registros) e 1/3 da base (288 registros) para teste. Foi mantida nos dois conjuntos a mesma proporção de exemplos para as duas classes: 61% dos alunos completaram o curso 39% não completaram. No conjunto de teste um ponto de interrogação foi colocado no lugar do valor a ser retornado como resultado da previsão. Os resultados referentes a acurácia dos classificadores não são diretamente obtidos da ferramenta necessitando de cálculos adicionais.

Resultados Experimento 3: A Tabela 3 mostra a acurácia dos classificadores para o conjunto de teste, ocorreu uma diferença maior comparando os resultados dos algoritmos JRip, DecisionTable, SimpleCart, MultilayerPerceptron e o *baseline* OneR. A Tabela 3 também mostra a matriz de confusão, composta pela classe positiva, alunos que concluíram, e negativa, alunos que não concluíram o curso e apresenta as taxas de acerto e erro dos classificadores.

Tabela 3. Acurácia e taxas dos classificadores especificando os conjuntos de treinamento e teste

Classificador	OR	JR	DT	SC	J48	RF	SL	MP	NB	BN
Acurácia	81,94	76,04	72,92	76,39	80,21	80,21	82,29	74,31	81,25	80,56
matriz de confusão	162 16 36 74	142 36 33 77	146 32 46 64	145 33 35 75	161 17 40 70	153 25 32 78	167 11 40 70	134 44 30 80	162 16 38 72	161 17 39 71
Verdadeiro Positivo	0,91	0,80	0,82	0,81	0,90	0,86	0,94	0,75	0,91	0,90
Falso Negativo	0,09	0,20	0,18	0,19	0,10	0,14	0,06	0,25	0,09	0,10
Verdadeiro Negativo	0,67	0,70	0,58	0,68	0,64	0,71	0,64	0,73	0,65	0,65
Falso Positivo	0,33	0,30	0,42	0,32	0,36	0,29	0,36	0,27	0,35	0,35

5.4. Discussão dos Resultados

Comparando-se os resultados dos três experimentos, verifica-se que a acurácia varia em média em torno de 75 a 80%. No entanto, um segundo nível de análise deve ser empreendido na comparação dos algoritmos. Além do valor da acurácia, um algoritmo pode diferir do outro nos valores das taxas de acerto e erro na classificação dos

exemplos positivos e negativos (Tabela 3). Um classificador que possui uma elevada taxa de erro para falso positivo não é adequado para a solução do problema, neste caso, considera-se um erro grave do algoritmo classificar um aluno com risco de evasão como sem risco. O erro do algoritmo de classificar um aluno no grupo de risco de evasão sem de fato ocorrer à evasão, falso negativo, é considerado um erro brando, menos grave. Observam-se que os algoritmos de classificação mais sofisticados, como MultilayerPerceptron e RandomForest apresentaram erro grave menor (taxa de erro falso positivo Tabela 3). O classificador DecisionTable apresentou pior resultado na acurácia e na taxa de erro do falso positivo, os demais apresentaram uma taxa de erro superior a 30%, portanto, tem-se que 30% dos alunos com risco de evasão não são corretamente classificados.

Outra análise feita para identificar as causas dos erros dos classificados foi realizada diretamente na base de dados. Observou-se que os dados (características) de alguns estudantes não seguiam o padrão das classes a qual eles pertenciam. Por exemplo, alunos com rendimento acadêmico a baixo da média, mas concluíram o curso, foram encontrados 67 registros (11,18%) no conjunto de treinamento e 25 registros (8,68%) no conjunto de teste. Outro grupo de alunos que possui comportamento fora do padrão da classe de alunos que evade, são os que possuem rendimento acadêmico elevado, mas não completaram o curso 75 registros (12,52%) no conjunto de treinamento e 31 registros (10,76%) no conjunto de testes. Este viés da base reflete na acurácia dos classificadores e reflete no aumento da taxa de erro, a remoção destes exemplos não é aconselhável porque a base perderia seu reflexo da realidade.

6. Conclusões e Trabalhos Futuros

A identificação dos alunos que apresentam risco de evasão através do uso técnicas de mineração de dados mostrou-se viável. Este trabalho avaliou a técnica através de três experimentos onde foram aplicados dez algoritmos de classificação sobre uma base de dados dos alunos de graduação do curso de Engenharia Civil da UFRJ. Os experimentos retornaram dados com acurácia média variando entre 75 a 80%.

Os desempenhos obtidos pelos algoritmos de mineração de dados dos mais simples aos mais sofisticados foram semelhantes. Além da acurácia, a taxa de erro dos classificadores foi considerada na análise, portanto, a previsão incorreta de alunos com risco de evasão é considerada erro grave do classificador. No entanto, a acurácia dos classificadores e a taxa de erro são fortemente influenciadas pelos vieses da base de dados, isto é, alunos que evadem do curso mesmo com rendimento acadêmico alto e alunos que concluem o curso com rendimento acadêmico abaixo da média, estes casos estão fora do padrão das classes aprendidas pelos classificadores.

O presente estudo ainda está em fase inicial, no entanto, já indica que é possível fazer a previsão de alunos com risco de evasão pode ser feita a partir de um número reduzido de atributos, por exemplo, verificou-se que o atributo mais importante é o coeficiente de rendimento do primeiro semestre letivo, o segundo é a nota na disciplina de Cálculo Diferencial e Integral I.

A qualidade dos resultados iniciais abre a possibilidade de novas investigações futuras, como por exemplo, o desenvolvimento de uma ferramenta de auxílio acadêmico que identifique quais alunos apresentam maiores riscos de abandonar os estudos de graduação. Os benefícios diretos da aplicação da mineração de dados neste contexto

são: i) identificar ainda nos primeiros semestres do curso os alunos mais propensos a evasão; ii) permitir que a universidade não utilize apenas dados estatísticos na análise do problema da evasão. A análise dos atributos permite identificar os fatores de sucesso e insucesso específicos para cada curso e relacionar estes fatores ao currículo do curso.

Como trabalhos futuros, consideramos aplicar procedimentos semelhantes para outros cursos da universidade, verificando se os resultados até agora observados se repetem para outros sub-conjuntos de alunos de graduação. Além das duas classes de alunos discutidas neste trabalho, outra classe de alunos deve ser analisada: os alunos que permanecem no curso além do prazo de conclusão médio.

Referências

Baker, R. and Yacef K. (2009) "The State of Educational Data Mining in 2009: A Review and Future Visions." Pages 3-17. JEDM -Journal of Educational Data Mining, 2009, Volume 1, Issue 1, October 2009

Baker, R., Isotani, S., Carvalho, A. (2011) Mineração de Dados Educacionais: Oportunidades para o Brasil. Revista Brasileira de Informática na Educação, 19(2), 3-13. http://dx.doi.org/10.5753/RBIE.2011.19.02.03

Barroso, M. F. e Falcão, E. B. M. (2004) "Evasão Universitária: O Caso do Instituto de Física da UFRJ", IX Encontro Nacional de Pesquisa em Ensino de Física.

Bouckaert R., Eibe F. Mark Hall, Kirkby, R., Reutemann, P., Seewald, A., and Scuse, D. (2010) "WEKA Manual for Version 3-6-4". December

Carvalho, L. A. V. (2005), Datamining – A mineração de Dados no Marketing, Medicina, Economia, Engenharia e Administração. Rio de Janeiro: Editora Ciência Moderna Ltda.

Castro F. et al. (2007) "Applying Data Mining Techniques to e-Learning Problems, Studies in Computational Intelligence (SCI)." 62, 183 - 221 (2007) Springer-Verlag Berlin Heidelberg.

Davies, P. (1997) "Within our control?: Improving retention rates" in FE, FEDA.

Dekker G., Pechenizkiy M. and Vleeshouwers J. (2009) "Predicting Students Drop Out: A Case Study". In Proceedings of the International Conference on Educational Data Mining, Cordoba, Spain, T. BARNES, M. DESMARAIS, C. ROMERO and S. VENTURA Eds., Pages 41-50.

Governo Federal (2007) "Diretrizes Gerais do Programa de Apoio a Planos de Reestruturação e Expansão das Universidades Federais – REUNI", http://portal.mec.gov.br/sesu/arquivos/pdf/diretrizesreuni.pdf, Fevereiro.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P. and Witten. I.H. (2009) "The WEKA Data Mining Software: An Update" SIGKDD Explorations, Volume 11, Issue 1.

Hämäläinen, W., Suhonen, J., Sutinen, E., and Toivonen, H. (2004) "Data mining in personalizing distance education courses". In world conference on open learning and distance education, Hong Kong, pp. 1–11

Han, J. (2005) "Feature selection based on rough set and information entropy". Granular Computing, IEEE International Conference on , vol.1, p. 153-158.

Han, J. and Kamber, M. (2006), Data Mining Concepts and Techniques, Morgan Kauffmann Publishers, Second Edition.

INEP - Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (2009) "Investimentos Públicos em Educação", http://portal.inep.gov.br/estatisticas-gastoseducacao e "Censo da Educação Superior', http://portal.inep.gov.br, Outubro.

Johnston V. (1997) "Why do first year students fail to progress to their second year? An academic staff perspective." Department of Mathematics, Napier University. Paper presented at the British Educational Research Association Annual Conference. September 11-14: University of York.

Kotsiantis, S., Pierrakeas, C. e Pintelas, P., (2003) "Preventing student dropout in distance learning using machine learning techniques." KES, eds. V. Palade, R. Howlett & L. Jain, Springer, volume 2774 of Lecture Notes in Computer Science, pp. 267–274

Minaei-Bidgoli, B., Tan, P., Kortemeyer G. e Punch, W.F. (2006) "Association analysis for a web-based educational system." Livro Data Mining in E-Learning. WitPress. Southampton, Boston.

Moore, R. (1995) "Retention rates research project" final report, Sheffield Hallam University.

Saraiva, S. e Masson. M. (2003) "Evasão e Permanência em uma Instituição de Tradição: um estudo sobre o processo de evasão de estudantes em cursos de Engenharia na Escola Politécnica da UFRJ", Relatório de Pesquisa.

Soares, I. S. (2006) "Evasão, retenção e orientação acadêmica: UFRJ – Engenharia de Produção – Estudo de Caso". Anais do XXXIV Congresso Brasileiro de Ensino de Engenharia - COBENGE. Passo Fundo: Ed. Universidade de Passo Fundo.