Identificação dos Fatores que Influenciam a Evasão em Cursos de Graduação Através de Sistemas Baseados em Mineração de Dados: Uma Abordagem Quantitativa

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

[1]Programa de Engenharia de Sistemas e Computação de Informática – Universidade Federal do Rio de Janeiro (UFRJ)
Caixa Postal 68.511 – 21941-972 – Rio de Janeiro – RJ – Brasil

[2]Departamento de Matemática – Universidade Federal Rural do Rio de Janeiro (UFRRJ)
Seropédica, RJ - Brasil

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

Abstract. *This paper uses data mining techniques to indentify key variables related with students failures in completing their undergraduate studies. In our approach, classification analysis is used to manipulate academic data of students of the largest Brazilian Federal University. Differently from other works, our research shows that even analyzing three different classes of students it was possible to have a global precision above 80%. The Naïve Bayes model was used to visualize the key variables used to separate the distinct classes of students.*

Resumo. *Este artigo utiliza técnicas de mineração de dados para identificar problemas relacionados com alunos que não conseguem completar os seus cursos de graduação. Nessa abordagem, a classificação manipula informações acadêmicas de alunos oriundos de uma grande Universidade Federal Brasileira. Muitas técnicas de mineração de dados foram avaliadas em função da acurácia obtida quando aplicadas a um conjunto de dados dos estudantes universitários. Os resultados demonstram que, mesmo analisando três diferentes classes de alunos, foi possível ter uma precisão global acima de 80%. O modelo Naive Bayes foi usado para visualizar os fatores de que distinguem os alunos que obtêm sucesso ou fracasso em seus cursos.*

1. Introdução

O sistema educacional brasileiro possui um grande número de estudantes que iniciam um curso universitário, mas não conseguem obter êxito em cumprir as exigências curriculares e se graduar. A evasão dos alunos que não completam o curso de graduação se configura como um dos grandes problemas que ocorre em instituições públicas e particulares [INEP 2009].

Dentro do contexto da educação pública superior, este tema foi abordado no Programa de Apoio a Planos de Reestruturação e Expansão das Universidades Federais

– REUNI, instituído pelo Decreto nº 6.096, de 24 de abril de 2007 [Governo Federal 2007]. Neste documento, o governo relata que: "os índices de evasão de estudantes nos cursos de graduação atingem, em alguns casos, níveis alarmantes". Além da ampliação do acesso às universidades, o governo tem como objetivo promover a permanência na educação superior. A meta global que o governo almeja é a elevação gradual da taxa de conclusão média dos cursos de graduação presenciais para noventa por cento.

Com o propósito de tomar medidas para evitar os baixos índices de conclusão dos cursos universitários das Instituições Federais de Ensino Superior (IFES) faz-se necessário identificar quais são os fatores que contribuem para o insucesso dos estudantes. No entanto, é primordial determinar quais técnicas são mais adequadas para identificar estes fatores.

Mineração de dados vem sendo empregada em um grande número de aplicações comercias e científicas, sua utilização mostrou-se viável para resolver diversos problemas relacionados à investigação de informações úteis em bases de dados. Sendo assim, este trabalho discute a utilização de mineração de dados dentro do contexto da investigação e extração de informações relevantes dos alunos de graduação da UFRJ (Universidade Federal do Rio de Janeiro). Deste modo, foi possível identificar os principais fatores comuns aos grupos de alunos que conseguem sucesso ou insucesso na conclusão do curso de graduação.

Este trabalho tem como objetivo identificar, apresentar e quantificar as variáveis que representam os principais fatores que influenciam a conclusão, a evasão e permanência além do tempo médio para concluir o curso de graduação. Definimos para fins de utilização da mineração de dados três classes distintas de alunos a serem pesquisadas: i) alunos que concluíram o curso; ii) alunos que interromperam o curso em algum período antes da formatura (evasão) e, iii) alunos que permaneceram matriculados além do prazo médio para conclusão do curso. Neste trabalho avaliamos seis algoritmos de mineração de dados aplicados a uma base de dados composta por exemplos de três classes de alunos. Os algoritmos de mineração de dados avaliados apresentaram acurácia entre 70% a 86%. Por último, o trabalho apresenta uma avaliação quantitativa baseada nos resultados do algoritmo de classificação *Naive Bayes*.

Este trabalho está organizado da seguinte maneira: Na Seção 2 serão apresentados trabalhos relacionados. Na seção 3 será apresentada a proposta de solução. Em seguida, na seção 4, apresentação de um estudo de caso. Na Seção 5 apresentação das considerações finais deste trabalho.

2. Trabalhos Relacionados

A evasão e o desempenho acadêmico dos alunos são assuntos importantes e foco de constantes debates no meio universitário. Entretanto, existem poucos sistemas que apóiam os gestores universitários na identificação precoce de alunos que apresentam riscos de evasão e retenção nas IFES. Atualmente, o estudo envolvendo o tema está restrito a identificação das causas através do uso de métodos estatísticos desacoplados de sistemas de informação, como relatado em Saraiva e Masson (2003), Soares (2006) e em estudos mais antigos [Moore 1995; Johnston 1997; Davies 1997]. No âmbito da UFRJ, Barroso e Falcão (2004) identificaram entre os alunos do Instituto de Física, os principais fatores que são obstáculos na conclusão dos cursos de graduação daquela

unidade: i) econômico - impossibilidade de permanecer no curso por questões sócio-econômicas; ii) vocacional – o aluno não se identificou 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.

Através da adoção de técnicas de mineração de dados ampliam-se as possibilidades de análise. Por exemplo, as investigações dos fatores de insucesso acadêmico podem ser feita sob novas perspectivas analíticas: a) investigação das diversas relações existentes entre os fatores que influenciam o insucesso dos alunos; b) identificação precoce dos alunos que estão sob os riscos de evasão ou que já apresentam indícios de serem influenciados pelos fatores de insucesso acadêmico.

A utilização de técnicas de mineração de dados sobre dados educacionais é relativamente recente conforme destacado por Baker et al. (2009, 2011) e Dekker et al. (2009). A maioria dos trabalhos correlatos restringe-se a identificar resultados em pequenos contextos relativos a apenas uma disciplina de curso não presencial e a análise de duas classes de alunos (sucesso e insucesso acadêmico).

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 demonstraram 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 online.

Hämäläinen et al. (2004) analisaram duas disciplinas de programação de computadores em um curso online. 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, identificaram-se 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, o classificador baseado em árvore de decisão foi considerado mais adequado.

A adoção de algoritmos de mineração de dados aplicados aos dados educacionais para a previsão da situação acadêmica dos alunos é um campo de investigação ainda não consolidado, necessitando de investigações complementares, tais como: definir quais são os atributos a serem utilizados e as técnicas de mineração de dados empregadas [Castro et al. 2007; Baker 2009; Dekker et al. 2009]. Estes autores indicam quais pontos precisam ser aprimorados na utilização de mineração de dados para identificação de estudantes com risco de evasão, a saber: i) transformação dos dados (os dados colhidos nem sempre são diretamente tratados pelos algoritmos de mineração); ii) identificar os atributos mais relevantes; iii) aplicar os algoritmos para identificar outros grupos (classes) de estudantes.

Neste trabalho, realizamos diversas transformações nos dados para aplicação dos algoritmos de mineração de dados como apresentadas na seção 3.2. O número de exemplos que compõe os subconjuntos da base de dados estudada é significativamente maior que os demais trabalhos encontrados na literatura (Tabela 1). Este trabalho obteve resultados elevados para acurácia dos algoritmos de classificação quando tratou três classes distintas de alunos, enquanto todos os outros trabalhos relacionados trataram apenas duas classes de dados. Outra diferença significativa deste trabalho com relação aos demais está na abrangência de sua aplicação, este trabalho foi aplicado aos cursos de graduação de uma grande universidade pública brasileira e os demais estão restritos a avaliação de uma disciplina ou curso.

3. Proposta de Solução

A abordagem adotada consiste em testar diversos algoritmos de mineração de dados e verificar sua precisão (acurácia) para identificar os subconjuntos (classes) de alunos que compõem a base de dados do sistema acadêmico da universidade. Os modelos gerados pelos algoritmos de mineração de dados apresentam um determinado nível de interpretabilidade, isso significa o quanto o modelo gerado pode ser compreendido ou interpretado pelo humano. Entre os algoritmos avaliados, o *Naive Bayes* foi escolhido, pois apresenta um modelo interpretável e seus resultados numéricos podem ser facilmente convertidos em gráficos para fazer a análise quantitativa dos principais fatores relacionada conclusão, evasão e permanência dos estudantes além do prazo médio para conclusão do curso. Para isto, foi utilizado um estudo de caso envolvendo os alunos de graduação da UFRJ. Neste estudo, as técnicas de mineração de dados foram utilizadas para identificar as relações existentes entre os fatos ocorridos durante a vida acadêmica dos alunos e relacioná-los com o sucesso ou insucesso em completar o curso de graduação.

Os algoritmos foram executados a partir da ferramenta de mineração de dados Weka, detalhes em Hall et al. (2009) e Bouckaert (2010), ela possui algoritmos de aprendizado de máquina que podem ser utilizados para extrair informações relevantes de uma base de dados. A ferramenta foi adotada, de acordo com os seguintes motivos: i) ferramenta *open-source* e livre de custos; ii) possuir várias versões de algoritmos empregados na mineração de dados; iii) disponibilidade de recursos estatísticos para comparar o desempenho dos algoritmos e apresentar diversos recursos para análise dos dados.

3.1. Descrição da Base de Dados

A base de dados utilizada neste estudo de caso é compartilhada com o Sistema Integrado de Gestão Acadêmica – SIGA, um sistema de gestão acadêmica que atende ao conjunto de alunos dessa universidade. Neste estudo de caso, foram selecionados dados acadêmicos dos alunos que ingressaram nos dois semestres letivos dos anos de 2003 e 2004. A base contempla alunos de 155 cursos de graduação oferecidos por 28 unidades da UFRJ.

3.2. Pré-processamento dos Dados

A etapa do pré-processamento dos dados, segundo Han e Kamber (2006) pertence ao Processo de Extração de Conhecimento – KDD, onde a mineração de dados é uma etapa

importante. Após a seleção dos dados e antes da aplicação das técnicas de mineração de dados, existe a etapa de seleção e transformação. Nesta etapa os atributos originais da base de dados são convertidos e/ou adaptados para aplicar a mineração de dados.

Escolheram-se os alunos que ingressaram nos anos de 2003 e 2004 porque estes já dispõem de situação acadêmica definida, previamente registrada no sistema acadêmico através do atributo "Situação Atual". No entanto, os dados registrados originalmente neste atributo possuem 22 valores diferentes. Portanto, para resumir e facilitar a análise dos dados foi introduzido o atributo "Status" que é uma redução e adaptação do atributo original "Situação Atual". O atributo "Status" foi utilizado como atributo identificador das três classes de alunos analisadas neste estudo de caso. O atributo possui três valores que descrevem a situação final da matrícula do aluno: cancelado, ativa e conclusão. O termo "cancelado" foi atribuído a todos os alunos com matrícula cancelada por: 1) iniciativa do aluno: cancelamento ou trancamento da matrícula; 2) iniciativa da instituição: matrícula cancelada por abandono do curso, não cumprimento das exigências curriculares e outros. O termo "ativa" foi atribuído a todos os alunos que tinham matrícula ativa e ultrapassaram o prazo médio para conclusão do curso. Por fim, o termo "conclusão" foi atribuído a todos os alunos que cumpriram com todos os requisitos da grade curricular do curso e foram diplomados.

A Tabela 1 ilustra os quatro conjuntos de alunos subdivididos nas três situações estabelecidas neste trabalho, as situações finais foram obtidas depois de 12 semestres letivos a partir do ano de ingresso (início do curso). Normalmente, os trabalhos relacionados na seção 2 abordam somente duas classes de alunos, neste estudo abordaremos o tratamento para três classes supracitadas.

Tabela 1. A Quantidade de Alunos por Situação Final

Ano de ingresso	Cancelado	Ativa	Conclusão	Total
2003-1	1342	715	1681	3738
2003-2	1124	696	1172	2992
2004-1	1597	1333	1259	4189
2004-2	1125	1424	556	3105

A Tabela 2 resume os atributos acadêmicos selecionados da base de dados do SIGA e suas respectivas adaptações a partir dos atributos originais. Além do atributo "Status" os atributos assinalados com (*) foram introduzidos para adaptação dos dados originais aos algoritmos de mineração de dados. O número total de atributos utilizados foi 75, isto envolveu dados acadêmicos de doze períodos letivos a partir do ano de ingresso no curso (2003-1).

Tabela 2. Atributos Acadêmicos dos Estudantes

Atributos	Descrição
Nome Curso	Nome do curso onde o aluno está matriculado
Descr. Unidade Curso	Nome da Escola, Instituto ou Faculdade onde o curso é oferecido na universidade
Situação Atual	Situação atual da matrícula do aluno
Status*	Atributo inserido, adaptação do atributo original "Situação Atual"
Ano Ingresso	Ano e período em que o aluno ingressou na universidade
CRA	Coeficiente de rendimento acadêmico acumulado (CRA): média de aproveitamento das disciplinas cursadas durante todo o curso de graduação
Período	Período letivo que cursou determinada disciplina

Atributos	Descrição
Situação Período	Situação da matrícula do aluno em cada período letivo
Nota	Nota da disciplina cursada no período
Situação Disciplina	Situação na disciplina: AP (Aprovado), RFM (Reprovado por Falta e Média), RM (Reprovado por Média) e RF (Reprovado por Falta)
NoP*	Atributo inserido guarda o número de disciplinas cursadas em cada período letivo
NoAP*	Atributo inserido guarda o número de disciplinas aprovadas em cada período letivo
MediaAP*	Atributo inserido guarda a média aritmética obtida nas disciplinas aprovadas em cada período letivo
NoRFM*	Atributo inserido guarda o número de disciplinas reprovadas por falta e/ou média em cada período letivo
NoRM*	Atributo inserido guarda o número de disciplinas reprovadas por média em cada período letivo

3.3. Aplicação dos Métodos de Classificação

Os métodos de classificação têm por finalidade encontrar um modelo que descreva uma classe ou conceito [Han e Kamber 2006]. O modelo obtido é utilizado para identificar novos exemplos cuja classe é desconhecida. Para criar um modelo, é necessário treinar os algoritmos de classificação utilizando exemplos corretamente rotulados em classes conhecidas *a priori*. Este processo denomina-se aprendizagem supervisionada. Neste caso, o algoritmo de classificação constrói o modelo (aprendizagem) a partir do conjunto de treinamento composto por amostras (exemplos) com a identificação da classe as quais elas pertencem. A aprendizagem é verificada quando o conjunto de teste é aplicado ao modelo aprendido. Existem várias formas de representar ou descrever um modelo, resultado dos algoritmos de classificação, as mais comuns são através de regras, árvores, tabelas de decisão, redes neurais, métodos estatísticos entre outros.

4. Estudo de Caso

Nesta seção, apresentamos os resultados obtidos após a aplicação dos algoritmos classificadores sobre a base de dados do sistema SIGA. Ressalta-se que nem todos os algoritmos implementados na ferramenta dão suporte a análise de três diferentes classes de exemplos. Portanto, a construção de modelos mais complexos envolve a escolha dos algoritmos que suportam análise multiclasse. Além disso, os algoritmos mais sofisticados demandam mais tempo para construir os modelos e os mais simples perdem um pouco na precisão dos modelos (acurácia), mas ganham na flexibilidade e interpretabilidade [Sumathi e Sivanandam 2006].

Existem vários métodos de divisão da base de dados para obter os subconjuntos de treinamento e teste para serem utilizados pelos algoritmos de classificação, um dos métodos mais empregados é a validação cruzada (*k-fold cross-validation*) [Han e Kamber 2006]. Neste estudo de caso optou-se por utilizar o método de validação cruzada com o número de conjuntos igual a 10 (k=10), devido ao grande número de exemplos disponíveis na base de dados. A Tabela 3 mostra as acurácias obtidas dos algoritmos de classificação utilizando três classes de alunos (ver seção 3.2) e o tempo de execução (em segundos) para construir o modelo.

As acurácias obtidas para os alunos de 2003-1 não variaram significativamente para as outras bases de dados 2003-2, 2004-1 e 2004-2 (ver Tabela 1). Optou-se, então, por apresentar apenas os resultados obtidos para a base de dados dos alunos que ingressaram no primeiro semestre de 2003-1. O tempo de execução apresentado na

Tabela 3 refere-se à construção do modelo para um dos dez conjuntos da validação cruzada. Observou-se um considerável aumento no tempo para a construção dos modelos para três classes de alunos, modelos simples utilizando duas classes de dados foram executados em poucos segundos. A construção do modelo do algoritmo *Multilayer Perceptron* levou mais de 12 horas de processamento.

Os algoritmos classificadores selecionados são os mais utilizados em mineração de dados segundo Wu et al. (2008) no artigo *Top 10 algorithms in data mining*. Acrescentou-se à Tabela 3 os resultados do algoritmo que utiliza redes neurais. Embora, o classificador *Naive Bayes* não tenha apresentado a melhor acurácia quando comparado aos demais algoritmos, seu rendimento global atende aos objetivos do trabalho. O modelo gerado pelo algoritmo é mais fácil de ser interpretado pelo ser humano e adaptado ao processo de visualização da informação [Han e Kamber 2006; Wu et al. 2008]. Tendo em vista o nosso objetivo de analisar os fatores e suas relações com o desempenho dos alunos no curso, adotou-se algoritmo classificador *Naive Bayes*. Os modelos gerados pelo algoritmo possibilitaram uma análise numérica e geração de gráficos facilitando a interpretação dos dados. A Tabela 3 mostra os resultados de diversos classificadores aplicados à base de dados estudada.

Tabela 3. Acurácia dos Classificadores e o Tempo em Segundos para Construção do Modelo

Método	Classificador	Acurácia	Tempo (segundos)
Árvore de decisão	*J48 (C4.5)*	83.0658	1.53
Vetor de suporte	*Support Vector machine (SVM)*	86.3563	82.95
Ensemble	*AdaBoost*	70.3585	2.04
Estatístico	*Naïve Bayes*	80.7116	0.3
Árvore de decisão	*SimpleCart*	82.932	24.57
Rede neurais	*MultilayerPerceptron*	85.2327	2317.79

4.1. Abordagem Quantitativa

Os gráficos da Figura 1 foram construídos a partir dos resultados obtidos com o classificador *Naive Bayes*. Os gráficos mostram no eixo x os semestres letivos cursados a partir de 2003-1, foram mostrados 12 semestres. No eixo y o número de disciplinas cursadas. Todos os gráficos contemplam as três classes de alunos (cancelado, ativa e conclusão), conforme mostra as legendas em cada gráfico. O Gráfico 1.a ilustra o número de disciplinas cursadas. O Gráfico 1.b apresenta o número de disciplinas aprovadas em cada semestre letivo. O Gráfico 1.c mostra o número de disciplinas onde os alunos foram reprovados por média (RM). O Gráfico 1.d ilustra o número de disciplinas onde os alunos foram reprovados por falta e média (RFM).

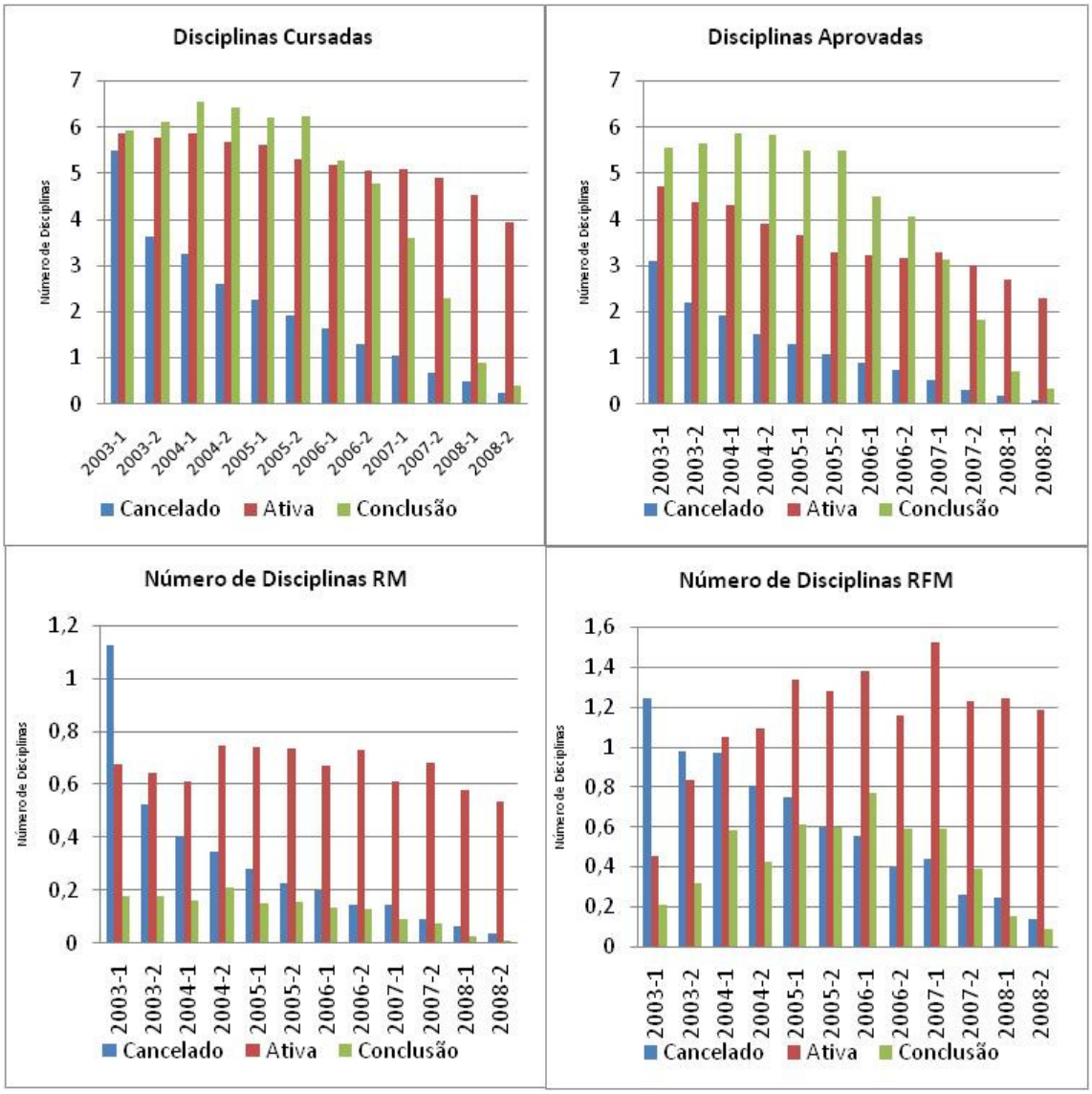

Figura 1. Da esquerda para direita temos: Gráfico 1.a Disciplinas Cursadas, Gráfico 1.b Disciplinas Aprovadas, Gráfico 1.c Número de Disciplinas RM e Gráfico 1.d Número de Disciplinas RFM.

A Figura 2 apresenta a média das notas das disciplinas aprovadas para cada classe de aluno. Observa-se que a média dos alunos cancelados cai abaixo do CRA (41,32) no primeiro ano acadêmico, diferente das outras duas classes que mantêm a média próxima do CRA durante o período de duração do curso (matrícula ativa CRA=62,12; conclusão CRA=78,17).

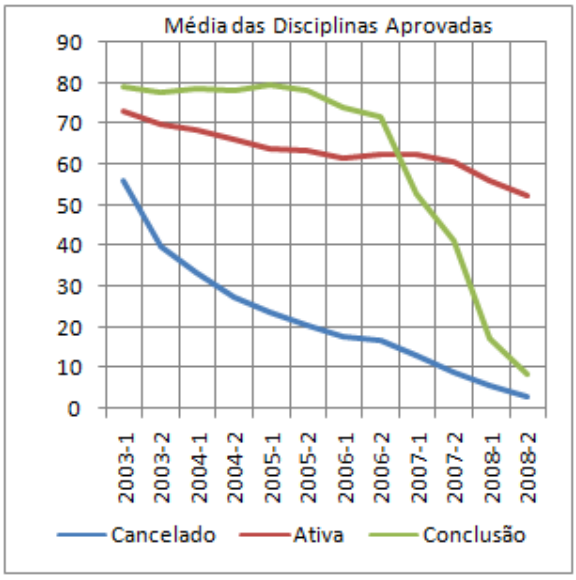

Figura 2. Média Aritmética das disciplinas cursadas e aprovadas

4.2. Validação dos Resultados

Procedimentos análogos foram realizados para os subconjuntos de alunos que ingressaram 2003 segundo semestre e 2004 primeiro e segundo semestres. Os resultados apresentaram valores bem próximos dos obtidos para base de 2003-1 (ver Figuras 1 e 2).

Tabela 4. Valores do primeiro semestre letivo do ano de 2004

Ano Ingresso	NoAlunos	SituaçãoFinal	CRA	NoDisCursadas	NoAP	MediaAP	NoFRM	NoRM
2004-1	1597	Cancelados	35,26	5,08	2,52	46,75	1,20	1,36
	1259	Conclusão	79,02	5,91	5,59	79,10	0,21	0,11
	1333	Ativa	66,21	5,73	4,95	74,31	0,32	0,46
2004-2	1125	Cancelados	35,07	5,04	2,54	47,33	0,96	1,54
	556	Conclusão	78,12	6,12	5,86	77,81	0,12	0,14
	1424	Ativa	68,97	5,95	5,31	74,93	0,25	0,38

A Tabela 4 mostra os valores numéricos para o ano de ingresso (2004-1 e 2004-2), número de alunos em cada situação final, CRA, número de disciplinas cursadas, número de disciplinas aprovadas, média de aprovação, número de disciplinas RFM e número de disciplinas RM. A análise dos valores da Tabela 4 deve ser comparada com os valores encontrados nos gráficos apresentados neste trabalho.

4.3. Análise dos Resultados

Como descrito anteriormente, a base de dados analisada contempla um número significativo de exemplos para cada subconjunto de alunos (Tabela 1). Por questão de espaço no artigo estão sendo apresentados resultados obtidos para o conjunto de alunos que ingressaram no primeiro semestre de 2003. No entanto, os dados dos alunos que ingressaram nos três semestres letivos consecutivos, 2003-2, 2004-1 e 2004-2 foram gerados e analisados seguindo os mesmos procedimentos descritos para base de 2003-1.

Optou-se por apresentar graficamente (Figuras 1 e 2) apenas um resultado (2003-1) e evitar fazer cálculo para obter a média dos resultados.

A seguir foram resumidas as informações mais relevantes sobre os fatores que caracterizam as três classes de alunos analisadas neste trabalho. Tais conclusões são baseadas nos dados obtidos em 2003-1 (Figuras 1 e 2) e nas análises comparativas dos três outros subconjuntos de dados citados acima.

Observa-se que os alunos com matrícula cancelada possuem as seguintes características: 1) possuem pelo menos uma disciplina RFM no primeiro período; 2) possuem pelo menos uma disciplina RM no primeiro período; 3) o número de disciplinas cursadas reduz ao longo dos períodos letivos; 4) o número de disciplinas aprovadas reduz a cada período, acompanhando o número de disciplinas cursadas (inscritas); 5) ao final do primeiro período, os alunos apresentam, nas disciplinas aprovadas, média inferior aos demais alunos; 6) ao final do segundo período, os alunos apresentam, nas disciplinas aprovadas, média próxima ou inferior ao valor do CRA dos alunos cancelados.

Com relação aos alunos que concluíram o curso, verifica-se que estes apresentam as seguintes características: 1) mantêm um número alto de disciplinas cursadas, igual ou superior a seis, diminuindo nos últimos períodos do curso; 2) possuem um alto índice de aprovações nas disciplinas; 3) mantêm a média das disciplinas aprovadas próximo do valor do CRA até o 8º período, após 8º período a média das disciplinas aprovadas vai diminuindo em relação ao CRA; 4) o número de disciplinas RFM destes alunos é maior do que os alunos cancelados a partir do meio do curso; 5) o número de disciplinas RM destes alunos é próximo de zero durante todo curso.

Os alunos que concluíram o curso possuíam uma regularidade de comportamento durante todo o curso, destacamos: o número elevado de disciplinas cursadas e médias altas de notas. Observa-se que, o número de reprovações aumenta ao final do curso, provavelmente em função do estágio curricular ou outra atividade.

Com relação ao subconjunto de alunos que mantiveram suas matrículas ativas até o ano/período de 2009-1, eles apresentaram as seguintes características: 1) possuem, ao longo do curso, um número elevado de disciplinas RFM com relação aos demais grupos de alunos; 2) possuem, ao longo do curso, um número elevado de disciplinas RM com relação aos demais grupos de alunos.

A partir das análises quantitativas, observa-se que os alunos com matrícula ativa são alunos que apresentam um comportamento regular ao longo de todo o curso, matriculam-se em um alto número de disciplinas, possuindo notas bem acima dos alunos que cancelaram, mas inferiores aos que concluíram. Além disso, possuem uma média constante de reprovações em disciplinas por falta ou por média.

5. Considerações Finais

Este artigo comparou seis algoritmos de classificação e apresentou uma abordagem quantitativa utilizando os resultados do algoritmo *Naive Bayes.* Este algoritmo apresenta um modelo interpretável e seus resultados numéricos podem ser facilmente convertidos em gráficos, ele obteve uma precisão de acerto em torno de 80% quando aplicado a base

de dados de informações acadêmicas dos alunos. Neste trabalho, a base de dados foi organizada para contemplar três classes distintas de alunos: alunos que concluíram o curso obtendo o diploma, não conseguiram concluir o curso e alunos possuíam matrícula ativa depois do prazo médio de conclusão do curso de graduação na IFES. Além das duas primeiras classes de alunos que são normalmente discutidas em outros trabalhos, o algoritmo *Naive Bayes* atendeu ao propósito de tratar a terceira classe de alunos.

A qualidade dos resultados obtidos abre a possibilidade de novas investigações, como por exemplo, o desenvolvimento de um sistema de informação capaz de auxiliar a gestão acadêmica das universidades. Os benefícios diretos da aplicação da mineração de dados neste contexto são: i) identificar ao longo do curso os alunos mais propensos a evasão e aqueles com possibilidade de permanecerem matriculados além do prazo médio para conclusão do curso; 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 períodos de cursos da universidade, verificando se os resultados até agora observados se repetem para outros subconjuntos de alunos de graduação e conceber um sistema de informação inteligente capaz de identificar alunos apresentam maiores riscos de não completarem os estudos de graduação.

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. e 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.

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. 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.

Sumathi, S. and Sivanandam, S.N. (2006). "Introduction to Data Mining and its Applications". Springer-Verlag Berlin Heidelberg 2006.

Wu, X., Kumar, V., Ross, Q.J., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G., Ng, A., Liu, B.,Yu, P., Zhou, Z., Steinbach, M., Hand, D. and Steinberg, D. (2008) Top 10 algorithms in data mining. Journal of Knowledge and Information Systems. Springer London. page 1-37. vol. 14, Issue 1.